UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA
Departamento de Filología Inglesa I

MODELOS DEL C-TEST Y SU CORRELACIÓN CON LOS TESTS DE LA COMPRESIÓN LECTORA DE LA E.O.I.

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Secundina Cantera Adrados

Bajo la dirección del doctor

Honesto Herrera Soler

Madrid, 2011

ISBN: 978-84-694-2880-1

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA

Departamento de Filología Inglesa I

MODELOS DEL C-TEST Y SU CORRELACIÓN CON LOS TESTS DE LA COMPRENSIÓN LECTORA DE LA E.O.I.

TESIS DOCTORAL

SECUNDINA CANTERA ADRADOS

Director: Dr. Honesto Herrera Soler

Madrid, 2010

ÍNDICE

Capítulo 4. EL C-TEST COMO PRUEBA DE COMPRENSIÓN LECTORA Y DE COMPETENCIA DE LA LENGUA

Capítulo 5. LA EVALUACIÓN EN LA ESCUELA OFICIAL DE IDIOMAS

Segunda Parte. INVESTIGACIÓN EMPÍRICA

Capítulo 6. PROCESO METODOLÓGICO

Capítulo 7. ANÁLISIS DE LA RECUPERACIÓN DE LOS TÉRMINOS LÉXICOS

Capítulo 8. ESTUDIO EMPÍRICO DEL C-TEST

AGRADECIMIENTOS

Hay muchas personas que han contribuido a que esta tesis llegara a buen puerto y a las cuales quisiera dar las gracias.

En primer lugar, me gustaría mostrar mi gratitud al Dr. D. Honesto Herrera Soler, director de la tesis, por todo el apoyo recibido, su paciencia, su disponibilidad, sus ánimos en épocas difíciles, su ayuda en la orientación y realización de todo el trabajo de investigación y sobre todo por haber confiado en mí más que yo misma.

En segundo lugar, quiero expresar mi agradecimiento a mis profesores del curso de Doctorado que supieron inculcar en mí su entusiasmo por la investigación, y me llevaron de la mano haciéndome ver los aciertos y los errores en mis primeros pasos por el apasionante mundo de la investigación. Doy las gracias también a la Dra. Gitte Kristiansen por su apoyo, sus consejos y por animarme a continuar con la tesis después del trabajo de investigación para la obtención del Diploma de Estudios Avanzados.

Esta tesis tampoco habría sido posible sin la colaboración de los profesores del departamento de inglés de la EOI Jesús Maestro de Madrid, que permitieron que impartiera clases a varios grupos de nivel intermedio facilitándome así la administración a dichos alumnos de varios tests para su pilotaje. Quiero expresar mi agradecimiento sincero a los profesores que aplicaron a sus alumnos la batería de tests de la EOI conjuntamente con el C-test y el cuestionario que se diseñó, siguiendo fielmente mis indicaciones, y facilitándome la tarea de recogida de datos en varia clases y a distintas horas. Gracias especialmente a Blanca Valle, Jefa del Departamento de inglés, a

Susan Thurgood, Francisco Bazaga, Fernando Alba, Ángela Baracaldo y Rocío Pérez por su ayuda desinteresada.

Quiero, así mismo, agradecer su participación a los alumnos que formaron parte del estudio empírico de esta tesis, cuyo último objetivo es la mejora de la batería de tests que se está utilizando para la evaluación de todos los alumnos de las Escuelas Oficiales de Idiomas de la Comunidad de Madrid. Estos alumnos son la razón por la que se ha realizado esta investigación.

También quiero expresar mi agradecimiento a todos mis amigos y personas cercanas que han tenido que soportar mis nervios, me han apoyado y han comprendido mi falta de tiempo para estar con ellos todo lo que yo hubiera deseado. He de hacer una mención especial a mi amigo Isidro Almendárez que dirigió mis pasos hacia la enseñanza del idioma inglés, y me puso en contacto con mi director de tesis.

Finalmente, agradezco a Beni y a Marcelo el haber estado ahí acompañándome y dándome ánimos durante todo este tiempo.

ABREVIATURAS

APA	American Psychological Association
APIEL	Advanced Placement International English Language
CALL	Computer Assisted Language Learning
CBLT	Computer Based Language Testing
CBT	Computer Based Test
CEFR	Common European Framework of Reference for Languages
CPE	Cambridge Proficiency Examination
CPH	Critical Period Hypothesis
DIALANG	Diagnostic Language Test
EAC	English through Academic Context
EAP	English for Academic Purposes
EFL	English as a Foreign Language
EOP	English for Occupational Purposes
ESOL	English for Speakers of other Languages
ESP	English for Specific Purposes
ETS	Educational Testing Service
GE	General English
GMAT	Graduate Management Admission Test
IELTS	International English Language Testing System
ILTA	International Language Testing Association
L1	First Language
L2	Second Language
LFP	Lexical Frequency Profile
LID	Local Item Dependence
LT	Language Testing
MELAB	Michigan English Language Assessment Battery

SLA	Second Language Acquisition
STEP	Special Test of English Proficiency
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication
UCLES	University of Cambridge Local Examinations Syndicate

INTRODUCCIÓN GENERAL

1. Motivación y objetivos de la tesis

Este trabajo de investigación está basado en la teoría de la Redundancia Reducida que a su vez proviene de la Teoría de la Información la cual establece que todo texto o comunicación es redundante, es decir contiene más información de la estrictamente necesaria para entender el mensaje, con lo cual si un texto o pasaje ha sido dañado se puede volver a reconstruir a partir de las partes que han sido transmitidas intactas.

Tanto el C-test como el *cloze* son pruebas integrales basadas en la teoría de la Redundancia de acuerdo con Klein-Braley y Raatz (1984), Klein-Braley (1985), Klein-Braley (1987), y Katona y Dörney (1993). Nuestro propósito fue explorar si el C-test era una alternativa válida al *cloze* que se venía utilizando en las Escuelas Oficiales de Idiomas y cuyos resultados no eran muy satisfactorios, ya que aún trabajando con un mismo texto, el grado de dificultad variaba enormemente dependiendo de la distancia entre los huecos o las palabras que los alumnos tuvieran que recuperar. También se quiso estudiar el efecto que la palabra que se tenía que restaurar en el C-test podría afectar a los resultados finales de la prueba.

2. Hipótesis, método y procedimiento de la investigación

Se partió de las siguientes hipótesis:

1. La palabra con la que se empieza a mutilar el texto no afecta a los resultados finales de los examinandos, es decir, no hay diferencias significativas en la puntuación final de los modelos de C-tests creados independientemente del punto donde se empiece a mutilar el texto que luego se ha de restaurar.

2. No existen diferencias significativas entre los dos modelos de C-test. Es decir, ambos modelos de C-tests creados son equivalentes.

3. Existen diferencias significativas entre las medias de los términos funcionales y los léxicos recuperados.

4. Existen diferencias significativas entre los super-ítems creados tanto a nivel léxico como a nivel funcional.

5. Las palabras pautadas se recuperan más fácilmente que las no pautadas.

6. Existen correlaciones significativas entre el C-test y el *cloze.*

7. Existe correlación entre el conjunto de pruebas de la EOI y el C-test.

8. Existe correlación entre los tests de la comprensión lectora de la EOI y el C-test.

Se seleccionaron 4 textos cortos, se ordenaron de menor a mayor dificultad y basándonos en esos mismos textos se diseñaron la dos únicas variantes posibles de C-test (C-test A y C-test B) cuya única diferencia era el punto donde se comenzaba a mutilar los textos.

La muestra fue de 151 alumnos de nivel intermedio pertenecientes a distintas aulas y a distintos horarios de la EOI Central de Madrid. Las dos variantes del C-test (C-test A y C-test B) se distribuyeron de forma aleatoria, dependiendo de donde se habían sentado los alumnos. Además, para evitar el efecto aprendizaje que se daría si los mismos alumnos repitieran las dos variantes del C-test, todos los examinandos hicieron un mismo ejercicio semejante al C-test en cuanto a estructura y grado de dificultad y se comprobó la homogeneidad de los grupos, ya que no había diferencias significativas entre las medias de las puntuaciones de los alumnos en esta prueba común.

Los 151 alumnos realizaron también una batería de tests de la EOI que medían su competencia general en la lengua inglesa y con los cuales se iba a estudiar los coeficientes de correlación de los C-tests. Así mismo, al final de los tests todos los alumnos que participaron en este estudio rellenaron un cuestionario que nos ayudará a conocer la opinión que tienen sobre este nuevo modelo de examen.

3. Principales resultados

El resultado final, después de haber comprobado la fiabilidad y la validez de los C-tests A y B, fue que existe correlación significativa y a veces muy significativa entre los dos modelos del C-test y el *cloze* y entre los dos modelos del C-test y el conjunto de pruebas actuales de la EOI. También se observó que existen diferencias significativas entre los resultados de los "super-items" o pasajes individuales que forman el C-test dependiendo de si se suprime la segunda o la tercera palabra. Sin embargo, estas diferencias quedan contrarrestadas y anuladas cuando consideramos los resultados del C-test en su conjunto (los 4 textos que forman el C-test), ya que en este caso no hay diferencias significativas entre las medias de la puntuación de los alumnos que realizaron el C-test A y el C-test B.

Se puede afirmar entonces que los dos C-tests creados son equivalentes, es decir, que el punto en el que se ha comenzado a mutilar las palabras de los textos que forman los dos modelos de C-test no ha afectado a la puntuación final de los alumnos, lo contrario que ocurre con el *cloze* donde el comienzo de supresión de las palabras afecta de manera muy significativa a los resultados finales de la prueba. También se comprobó que los ítems funcionales se recuperan más fácilmente que los léxicos, y que por lo tanto existen diferencias significativas entre los subtests léxicos y funcionales creados a partir de los mismos textos.

4. Organización y contenido de la tesis

En *el primer capítulo* se analiza la situación actual de la evaluación en el estudio de idiomas. Se observan los distintos tipos de tests que existen en este momento y cual es su uso. Se estudia la evolución de los tests a lo largo de la historia y vemos la importancia que tienen en nuestra sociedad puesto que todos los ciudadanos se ven positiva o negativamente afectados por los mismos a lo largo de su vida. También exponemos las características que determinan que un test sea fácil o difícil, así como los factores que afectan a los resultados y a su interpretación. Por último se hace una predicción de cómo esperamos que los tests evolucionen en un futuro.

En el *segundo capítulo* estudiamos las principales características de los tests. Además, se expone la base teórica y los criterios que hay que seguir para elaborar un buen test que sea válido, fiable y que nos permita evaluar la habilidad que nos interesa y de este modo, partiendo de los datos obtenidos, poder generalizar los resultados y determinar la competencia de un candidato en una determinada destreza en una situación del mundo real.

Como nuestro interés principal es encontrar una alternativa al *cloze* que forma parte de la batería de pruebas usadas en la EOI para evaluar la habilidad de la comprensión lectora, en el *tercer capítulo* exponemos cual es la situación actual de la evaluación de la comprensión lectora, de qué depende la eficiencia de un lector a la hora de comprender un texto y las distintas estrategias utilizadas en el proceso. Así mismo, se analizan las diferentes teorías sobre si la comprensión lectora es una habilidad que se debe medir de forma global por ser un proceso unitario o por el contrario, se compone de varias destrezas que se pueden medir independientemente. Estudiamos también los distintos factores que pueden afectar a la comprensión de un texto, centrándonos especialmente en la importancia que el vocabulario tiene para la comprensión lectora, y terminamos el capítulo enumerando algunas de las técnicas o

métodos existentes en este momento para medir la habilidad de comprensión lectora de un individuo.

En el *capítulo cuarto* se hace un estudio más profundo de la técnica del C-test como prueba de redundancia reducida. Se exponen las características de este test enumerando las ventajas y desventajas que presenta con relación al *cloze* que es el test al que intentaba mejorar y sustituir. Se analizan las investigaciones más importantes que se han realizado hasta el momento sobre el C-test y se intenta definir que es lo que mide este test, de acuerdo con los distintos trabajos publicados al respecto. También exponemos los conocimientos que se requieren para recuperar una palabra que ha sido mutilada. Finalmente, se señalan los principales usos del C-test y se enumeran algunos factores que hay que tener en cuenta a la hora de elaborarlo.

En el *capítulo quinto* se muestra el origen de la EOI, la importancia de su existencia para la enseñanza de idiomas de las personas adultas y la evolución que ha sufrido la evaluación de idiomas en la Escuela Central de Idiomas de Madrid en los últimos años para crear exámenes más válidos y fiables y también para adaptarse al Marco Común Europeo de Referencia para las Lenguas.

Los objetivos de la tesis, los materiales, las hipótesis, el método y el procedimiento seguidos se describen en el *capítulo sexto*.

En el *capítulo séptimo* se estudia la recuperación de los términos léxicos de los dos modelos de C-test y se analizan los factores que influyen en que una palabra se recupere más o menos fácilmente.

En el *capítulo octavo* se realiza el estudio empírico de todos los datos obtenidos en el trabajo de investigación. Se construyen dos modelos de C-test variando el punto donde se empiezan a mutilar las palabras. Éstas se empiezan a mutilar en la segunda o en la tercera palabra después del primer punto de cada texto o subtest. Se pautan las palabras mutiladas en la mitad de los textos de cada modelo de C-test para estudiar si este factor influye en la

recuperación de los términos. Cada modelo de C-test lo hace un grupo de alumnos diferente, por lo que se estudia si los dos grupos de alumnos que forman parte de la investigación, grupo A y grupo B, pueden considerase homogéneos. Esto se realiza a través de un test al que se le ha llamado Test del Tutor. Los resultados nos indican que no existen diferencias significativas entre las medias de los componentes de los dos grupos, por lo que se puede considerar que existe homogeneidad entre los mismos. Se hace una división entre los términos léxicos y los funcionales y se comprueba que los términos funcionales se recuperan más fácilmente que los términos léxicos en los dos modelos de C-test. Igualmente se comprueba que los términos pautados se recuperan más fácilmente que los no pautados.

También se demuestra que existen diferencias significativas entre las medias de todos y cada uno de los subtests que forman los dos modelos de C-test. Sin embargo, esas diferencias se neutralizan entre sí al considerar el C-test A y el C-test B en su conjunto. Los datos obtenidos nos indican que no existen diferencias significativas entre los resultados de los dos modelos de C-test, por lo que se pueden considerar tests paralelos o equivalentes. Esto quiere decir que el punto donde se ha empezado a mutilar las palabras al elaborar los dos modelos de C-test no ha afectado a los resultados. Por lo tanto, existe una clara ventaja en la elaboración del C-test respecto a la del *cloze* en el que según las palabras que se supriman los resultados son distintos.

Puesto que nuestro objetivo inicial era saber si era factible y apropiado sustituir el *cloze* que se está utilizando actualmente en la batería de exámenes de la EOI por el C-test, se decidió estudiar las correlaciones que existen entre ellos y ver si la diferencia de medias entre ambos es o no significativa.

Se estudia la validez y fiabilidad del C-test y de los tests de la EOI. Para ello se calcula la consistencia interna entre los componentes del C-test y la de los tests que forman el test de la EOI a través del Alfa de Cronbach y vemos que en ambos casos es alta, lo cual demuestra la fiabilidad de los tests. Así mismo, se analizan las correlaciones existentes entre todos los modelos de

C-test y los tests de la EOI así como las correlaciones entre los distintos componentes de cada test, encontrando que, en general, las correlaciones son significativas. De la misma forma, se estudia cual es el test que contribuye más y el que contribuye menos al test global de la EOI y también cual es la contribución de cada subtest al C-test.

En el *capítulo noveno* se analiza el cuestionario que se administró a los alumnos para estudiar una serie de factores, tales como la motivación, la actitud de los alumnos hacia el aprendizaje del inglés, la validez aparente del C-test, la influencia de la edad o del género en los resultados finales del C-test y la opinión de los participantes tanto sobre los conocimientos que el C-test puede medir, como si consideran que el C-test podría entrar a formar parte de la batería de exámenes de la EOI. Se analiza también si hay alguna relación o diferencia entre algunos de los factores del cuestionario y los resultados del C-test.

Finalmente, en el capítulo décimo se expresan las conclusiones a las que se ha llegado en esta tesis y se adelantan los planes para futuras investigaciones.

Primera Parte

FUNDAMENTOS TEÓRICOS

Capítulo 1

EVALUACIÓN DE IDIOMAS

1.1. Introducción

Hay muchas razones para evaluar un idioma en nuestra vida diaria. Los exámenes responden a la necesidad de medir la habilidad de las personas, de forma que podamos tomar medidas sobre su futura educación o empleo. Alderson (2000a: 27) afirma que la evaluación es valorada por la sociedad en si misma, y está por lo tanto justificada. Para McNamara los exámenes de idiomas se están enfrentando a importantes retos como resultado de nuestro mayor entendimiento del carácter social de sus constructos y sus prácticas.

> An awareness of language use as a social activity, of the socially derived nature of our notions of language, and of testing as an institutional practice, is causing language testers to look critically at their practices and the assumptions that underpin them. (McNamara, 2001b: 333)

Aparte de la preocupación por las consecuencias sociales de los exámenes está la idea de que el papel de los exámenes hay que redefinirlo para que sirva de mayor ayuda a los profesores y a los alumnos. Quizás haya que volver a lo que se llamó evaluación alternativa e interpretar su finalidad de forma bastante diferente.

En muchas escuelas de Estados Unidos han restringido los exámenes estandarizados de elección múltiple a favor de exámenes basados en una actuación más compleja de los candidatos. Muchos de los artículos sobre lingüística aplicada critican, desde una perspectiva social y posmoderna, que la mayoría de la investigación sobre la evaluación se realiza sobre exámenes a gran escala, ignorando los contextos de las clases donde se lleva a cabo la enseñanza y el aprendizaje. Así mismo, critican el uso de métodos cuantitativos, sofisticados técnicamente, para mejorar la calidad de los tests a expensas de otros métodos más accesibles para los no expertos. Sin embargo, no conviene olvidar el carácter operativo, económico, y practico de los tests de

elección múltiple en exámenes a gran escala cuando se trata de la selección de personal.

La tendencia y el reto actual de los investigadores, Shohamy (2001a; 2001b), McNamara (2001b), Messik (1989), es facilitar exámenes que sean más fáciles de usar dentro de una clase así como reflexionar sobre el papel de los exámenes como práctica social, sin eludir la responsabilidad que tienen frente a los administradores, los directivos, los alumnos y los profesores de idiomas. Las inmensas contribuciones de Bachman (1990) y Bachman y Palmer (1996) a la evaluación de idiomas, se basan en la teoría de Messick (1989) sobre la naturaleza fundamentalmente social de la validez de un test. Esto se ve reflejado en la siguiente figura que expresa la teoría unificada de la validez y que todavía tiene vigencia:

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASES	Contruct validity	Constuct validity+ Relevance/utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

Figura 1. Facetas de la validez. (Fuente: Messick, 1989:20; en McNamara, 2001b: 335)

"Construct validity" se refiere a la necesidad de recoger evidencia que apoye las interpretaciones que hacemos de los resultados, en función de los constructos que hemos propuesto.

“Construct validity + relevance / utility” resalta la necesidad de que los constructos sean relevantes y útiles en el contexto del examen.

“Value implications” insiste en que todas las interpretaciones de los resultados de los tests implican cuestiones de valor. Es decir, que para esta actividad, no disponemos de una base que sea objetiva, científica, y en la que no haya que valorar.

Finalmente, “social consequences” enfatiza la necesidad de investigar el impacto del examen. Numerosas investigaciones se están ocupando ahora de este tema.

Este trabajo de investigación tiene como meta buscar instrumentos de evaluación alternativos que sean no sólo económicos en cada etapa del proceso de diseño, construcción, administración y corrección sino también fiables y válidos. Para ello, analizaremos el marco teórico de la evaluación en general y de la técnica del C-test en particular para estudiar si ella pudiera ser una alternativa válida al *cloze test* que en este momento estamos usando en la EOI de Jesús Maestro.

1.2. Evaluation, test, assessment y measurement

Los términos *evaluation*, *test*, *assessment*, y *measurement* (evaluación, prueba, valoración y medición) a veces se usan como sinónimos para referirse a algunas actividades profesionales y aunque son superficialmente similares también tienen sus propias características. Huitt et al. consideran que *assessment, measurement, research*, y *evaluation* son procesos distintos, aunque algunos aspectos de los mismos estén solapados:

> Assessment refers to the collection of data to describe or better understand an issue, measurement is the process of quantifying assessment data, research refers to the use of data for the

> purpose of describing, predicting, and controlling as a means toward better understanding the phenomena under consideration, and evaluation refers to the comparison of data to a standard for the purpose of judging worth or quality. (Huitt et al., 2001)

Para Kizlik (2007), excluyendo unas pocas excepciones *measurement* se refiere al proceso por el cual se determinan las dimensiones físicas o los atributos de un objeto. En este proceso no se valora nada, simplemente se recoge información relacionada con alguna regla establecida:

> Measurement refers to the process by which the attributes or dimensions of some physical object are determined. …. We are not assessing anything; we are simply collecting information relative to some established rule or standard.

Una excepción parece ser el uso de la palabra *measure* para determinar el grado de inteligencia de una persona. A menudo se oye la frase, "this test measures IQ". También se puede utilizar para medir actitudes y preferencias. Pero en todos estos casos utilizamos el verbo *measure* con el significado de aplicar una medida estándar a un objeto, acontecimiento o condición siguiendo las prácticas previamente aceptadas por los expertos:

> To apply a standard scale or measuring device to an object, series of objects, events, or conditions, according to practices accepted by those who are skilled in the use of the device or scale.

De forma que cuando medimos, generalmente usamos algún instrumento estándar, mientras que según Kizlik (2007) *assessment* es el proceso por el que se obtiene información con alguna finalidad conocida. Un test es una forma especial de *assessment*:

> Assessment is a process by which information is obtained relative to some known objective or goal. Assessment is a broad term that includes testing. A test is a special form of assessment.

Bachman considera que *measurement* es un sinónimo de *assessment* igual que *examination* es un sinónimo de *test*. Ambos están de acuerdo, sin embargo, en que tanto v*aloración* como *medición* son términos amplios que incluyen la idea de *testing*. Considera que *meassurement* en ciencias sociales es el procedimiento de cuantificar las características de las personas siguiendo unos procedimientos explícitos:

> Measurement in social sciences is the process of quantifying the characteristics of persons according to explicit procedures and rules. (Bachman, 1990: 18-19)

En un sentido amplio, en las Ciencias Sociales, *mesurement* no se limita a los atributos de las personas. Lo mismo se puede medir el grado de consenso sobre una política de idioma como la frecuencia de una característica sintáctica en el discurso hablado. Según Bachman (1990), una de las características que diferencia *measurement* de *testing* es que en *testing* se cuantifican atributos o habilidades mentales que sólo pueden ser observados indirectamente por lo que recogemos números que después hay que analizarlos e interpretarlos.

> Physical attributes such as height and weight can be observed directly. In testing, however, we quantify mental attributes and abilities, sometimes called traits or constructs, which can only be observed indirectly. (Bachman, 1990: 19)

Puesto que nuestras observaciones de la actuación de un candidato son indirectas, y por lo tanto incompletas, imprecisas, subjetivas, y relativas, nuestras interpretaciones de los resultados de los tests también son limitadas.

Broadfoot piensa que *assessment* implica la intervención humana que valora e influye en cada etapa del proceso (diseño, elaboración, administración, etc.), por lo que no puede ser lo mismo que *medida*, que es una ciencia objetiva:

> Even the most apparently objective assessment - a multiple choice- is objective only in its scoring; it is not an objective assessment as such simply because all assessment involves professional judgement. (Broadfoot, 2005)

Un *test* puede ser considerado un instrumento de medida diseñado para obtener una muestra específica de la habilidad o comportamiento de un individuo. Como cualquier tipo de medida, el test cuantifica las características de los individuos de acuerdo con procedimientos explícitos. Lo que distingue a los tests de cualquier otro tipo de medida es que están diseñados para obtener una muestra específica de comportamiento. Si pudiéramos medir una habilidad del idioma basándonos en cualquier muestra de lenguaje que tomáramos, los tests no serían necesarios. Necesitamos los tests de idiomas para obtener la muestra que nos permita sacar conclusiones sobre una determinada habilidad. Los tests, por lo tanto, son los medios que nos aseguran que la muestra de lenguaje obtenida es suficiente para medir la habilidad deseada. Brown y Rodgers (2002: 33) también consideran que los tests son instrumentos encargados de recoger y recopilar los datos para "establecer el valor de algo con un propósito determinado".

Evaluación no implica necesariamente t*esting*. Cuando evaluamos, recogemos información que nos ayudará a hacer juicios acerca de ciertas situaciones. Cuanto más fiable y relevante sea la información que tengamos, más probabilidades tendremos de tomar la decisión correcta. Para evaluar no necesitamos obligatoriamente un test. Si tenemos, por ejemplo, que contratar a una persona, podemos tomar una decisión evaluando a la misma. Para ello podemos basarnos en informaciones verbales de alguna persona en cuyo criterio confiemos, en cartas de referencia, en la impresión general que nos ha causado, en el perfil y experiencia profesional que posee y por supuesto también en su expediente académico.

De la misma forma, un test no tiene por qué ser una evaluación. A veces se utilizan tests con fines pedagógicos como puede ser para repasar la materia

que ya se ha enseñado. En estos casos no se hace ninguna evaluación de los resultados obtenidos en los tests.

Bachman defiende que la evaluación tiene lugar solamente cuando los resultados de los tests se usan para tomar una decisión:

> It is only when the results of tests are used as a basis for making a decision that evaluation is involved. ….
> In summary, then not all measures are test, not all tests are evaluative, and not all evaluation involves either measurement or tests. (Bachman, 1990: 23)

Para Murphy (1985), la evaluación se ocupa de emitir juicios en el campo de la educación y ha adquirido prominencia en los últimos años, principalmente debido a la presión exterior sobre los educadores para que expliquen y justifiquen lo que están haciendo. También opina que, en la evaluación, los tests tienen una función útil pero limitada y que debemos tener cuidado con lo que pedimos o esperamos de los tests.

1.3. Importancia de los tests en nuestra sociedad

Los exámenes han adquirido un papel muy decisivo en nuestra sociedad, ya que forman parte de contextos políticos, sociales, educativos y económicos. Son instrumentos poderosos que afectan a personas concretas de nuestra sociedad. Los que crean los tests afirman que el criterio principal a la hora de confeccionar un buen test es la precisión, la fiabilidad o la validez, y una vez que han elaborado un test se olvidan de cómo es usado. Sin embargo, en realidad los tests se utilizan como armas poderosas para imponer nuevas políticas, para castigar, para excluir o para perpetuar poderes existentes.

Según Shohamy (2001a, 2001b), las autoridades educativas aseguran que la introducción de un nuevo examen se debe exclusivamente a motivos

educativos, es decir, para mejorar el conocimiento de los alumnos, pero muy frecuentemente los tests son usados para conseguir otros objetivos concretos.

Una de las limitaciones de los tests es que su constructo es muy limitado y aunque la fiabilidad y la validez de un examen intentan paliar en parte esa limitación, sin embargo, no dejamos de basarnos en una parte muy pequeña del lenguaje para recoger evidencia e inferir lo que no puede ser observado. Otra limitación puede ser la arbitrariedad de los valores que admitimos como apropiados cuando se decide la validez de un test. Tenemos que vigilar constantemente todos los pasos que se siguen para diseñar un ejercicio, para fijar su validez y fiabilidad, su administración, su corrección y el análisis de sus resultados y finalmente, cómo estos resultados son interpretados y utilizados, es decir, vigilar lo que se ha dado en llamar “consequential validity”. Finalmente, otra limitación que puede tener el test es que la consideración de la relevancia de algunos aspectos del test tiene un carácter subjetivo, ya que a la hora de evaluar tomamos decisiones personales sobre la información que vamos a ignorar y sobre los aspectos que se van a considerar como algo importante y representativo del conocimiento de una lengua.

Aunque la fiabilidad y validez de los correctores han mejorado considerablemente a la hora de evaluar, ya que se utilizan tablas cuando se corrigen pruebas subjetivas y por lo general se realizan cursos de estandarización dentro de cada institución, sin embargo, se ha hecho poco a la hora de comparar resultados con otros correctores de otras instituciones u otros contextos sociales diferentes o más generales. Shohamy hace una crítica a los tests tradicionales que no tienen en cuenta los motivos por los que se usan y las consecuencias de los mismos sobre las personas y sobre la sociedad.

> Traditional testing is not interested in the motives for introducing tests, in the intentions and rationale for using tests or in the examinations of whether intentions were fulfilled. It is not interested in the steps taken in preparation for tests or in how test takers feel about tests. It is especially not interested in the

> consequences of tests and their effects on those who failed or succeeded in them. It also overlooks how the test affected knowledge, learning patterns and habits. Traditional testing views tests as isolated events, detached from people, society, motives, intentions, uses, impacts, effects and consequences.
> (Shohamy, 2001a: 4)

Tenemos que prestar atención a la intención y a las razones por las que las autoridades educativas quieren introducir un test nuevo y a su impacto sobre los que se examinan, sobre los profesores que preparan a los alumnos para el test y sobre los conocimientos que se adquieren al preparar esos tests, es decir, sobre la enseñanza y aprendizaje del idioma y sobre las consecuencias que esos tests tienen en los sistemas educativos y la sociedad. Tenemos que analizar también si los tests son justos y éticos, ya que a veces los Gobiernos y las Empresas Comerciales los utilizan para una finalidad para la que no fueron creados.

De acuerdo con Messick, para que un test sea válido tenemos que prestar atención a todos los aspectos anteriormente citados, es decir, debemos procurar que los tests no sean sesgados en la corrección o en la interpretación de los resultados y debemos también tener en cuenta las consecuencias sociales de los mismos:

> The consequential aspect of construct validity includes evidence and rationale for evaluating the intended and unintended consequences of score interpretation and use in both the short and long term, especially those associated with bias in scoring and interpretation, with unfairness in test use, and with positive or negative washback effects on teaching and learning.
> (Messick, 1994: 251)

Hoy en día los tests están experimentando un gran cambio y se ha pasado de analizarlos desde un punto de vista puramente técnico a tener también en cuenta su uso y sus consecuencias. Los tests deciden y tienen un papel muy importante en nuestras vidas. Un examen decide si podemos ir a la

universidad o no, que carrera podemos estudiar, en que nivel podemos hacerlo, que certificados, premios o becas podemos conseguir, si podemos o no emigrar a un país para empezar una nueva vida, que trabajo podemos obtener, nuestros ascensos, nuestro salario y hasta nuestro reconocimiento social. Es decir, si la sociedad y nosotros mismos nos considerarnos fracasados o triunfadores. Shohamy lo describe muy bien cuando dice que los tests deciden a veces nuestra vida de forma irreversible:

> Tests, then, can open or close doors, provide or take away opportunities, and in general shape the lives of individuals in many different areas. It is often the performance on a single test, often on a single occasion at a single point in time, which can lead to irreversible, far-reaching and high-stake decision.
> (Shohamy, 2001a: 16)

También tienen un papel decisivo en muchos países, muchas empresas, y muchas universidades, que son valoradas de acuerdo con los resultados obtenidos por los alumnos en ciertos exámenes, aunque a veces los alumnos aprendan técnicas de examen para aprobar un test que no tienen nada que ver con los conocimientos adquiridos. Los exámenes, por lo tanto, en manos de muchos burócratas, pueden servir no para medir el conocimiento de un idioma sino como herramientas disciplinarias para diferenciar y juzgar a la gente, para excluir grupos, para limitar ayudas institucionales o como barreras que evitan la entrada de grupos no deseados por algunos gobiernos e instituciones.

En resumen, los exámenes pueden ser un mecanismo de poder y control que ha arraigado firmemente en los Negocios, Administraciones y Educación. Algunos tests se han convertido en instituciones por si mismos, sin ser cuestionados en absoluto a pesar del papel decisivo que juegan en el mercado laboral y en los planes educativos donde a veces son utilizados por los políticos para llevar a cabo reformas educativas.

Los que toman decisiones se apoyan en los tests para ganar credibilidad, ya que los tests disfrutan del prestigio de una ciencia al ser

herramientas vistas como objetivas, justas, verdaderas, fidedignas, y a pesar de que sabemos que la información dada por un test sobre una institución o una persona es ínfima y de que además de la validez y la fiabilidad de un test hay otros muchos factores que son importantes, tales como, el constructo, el método, cómo y cuándo se va a realizar el examen, cómo se va a puntuar y cómo se van a interpretar los resultados. Variables todas ellas que son decididas por las personas que ponen el examen o por las que toman decisiones que a veces utilizan los tests para resolver sus problemas. Hawthorne nos ilustra este problema al describirnos el uso del STEP (Special Test of English Proficiency) por las autoridades australianas como un instrumento para controlar la gran cantidad de personas que solicitaban asilo en el país.

> Designed to appear as a gatekeeping device, STEP was in fact constructed to facilitate the Australian government's acceptance of a large majority of asylum seekers – imposing a judicious appearance of control over what was potentially an unmanageable, expensive and diplomatically awkward situation. …….. Macro-political issues have a profound potential on test design, administration and outcomes. …… In cases such as the above, where the measurement of language proficiency is clearly a pretext for achieving some broader political purpose, construct validation procedures, which are concerned with what a test purports to measure, may be an insufficient means of ensuring a test's ethicality. (Hawthorne, 1997: 253-258)

Otro ejemplo es el de Letonia, que utilizó un test de idioma, cuando se independizó de la Unión Soviética en 1991, para restringir la ciudadanía a las personas de etnia lituana y a los que vivían allí antes de la anexión. Es decir, a los que podían hablar su idioma. Esto contribuyó a que muchos rusos que habían vivido en Letonia toda su vida decidieran abandonar su territorio. El número de rusos en este país descendió drásticamente desde un 52 por ciento a un 32 por ciento de la población total. Contra el poder de los que elaboran y administran los tests está el individuo que se somete a ellos y que no posee ningún dato para oponerse con alguna garantía a la industria e instituciones

que controlan los tests. Hanson se refiere a esta diferencia de poder y de actuación entre las instituciones y los examinandos.

> In nearly all cases tests givers are organizations, while test takers are individuals. Test-giving agencies use tests for the purpose of making decisions or taking actions with reference to test takers – if there are to pass a course, receive a certificate, be admitted to college, receive a fellowship, get a job or promotion. That, together with the fact that organizations are more powerful than individuals, means that the testing situation nearly always places test givers in a position of power over test takers. (Hanson, 1993: 19)

1.4. Historia de los tests

Los tests se originaron como herramientas para medir el conocimiento, principalmente con motivos de selección. Según documenta Spolsky (1995, 1997), los chinos utilizaron los tests en el año 210 antes de Cristo como poder político, social y educativo.

A mediados del siglo XIX la sociedad empezó a considerar los tests como una forma válida para seleccionar y controlar las escuelas. En Inglaterra, por ejemplo, los tests estaban ya asentados para finales del siglo XIX como método para evaluar el progreso de los alumnos de enseñanza primaria y secundaria. En Estados Unidos el auge de los tests se produjo en los años 1920 cuando se utilizaron por el ejército para detectar los reclutas que no eran adecuados para sus fines. Alrededor de esos años se produjo el cambio en la forma de evaluar lo que el alumno sabía o había conseguido aprender. Se empezó también a evaluar su conocimiento con relación a lo que sabía un grupo determinado. Esto ayudó a construir normas nacionales tanto para los

tests de inteligencia como para los tests de rendimiento, extendiéndose el movimiento de estandarización en la educación.

En el siglo XX, después de la Primera Guerra Mundial, aumentó la clase media y con ella la necesidad de selección. También se introdujo la educación para todos los ciudadanos que vino acompañada por la generalización de los tests para abrir las oportunidades a todos los alumnos, evitando así seleccionarlos basándose exclusivamente en su clase social o en recomendaciones.

Además de garantizar las mismas oportunidades a todos los alumnos, los tests debían ser objetivos, con lo que no se debía proporcionar ningún dato que pudiera aumentar la subjetividad o afectar a la discriminación de los que tomaban las decisiones. También debían ser válidos y fiables, con lo que ganaron la confianza de la sociedad que los consideraron como un método científico. Finalmente, los tests tendrían que ser también objetivos a la hora de ser corregidos, con lo que hubo que modificar, en la medida de lo posible, la subjetividad relacionada con la corrección y con los correctores. Más tarde cobraron importancia otros factores como el impacto y la autenticidad de las técnicas de examen. Actualmente se está investigando sobre los distintos procesos que tienen lugar cuando hay que responder a la tarea propuesta en un examen. Ello nos dará información para saber si estamos midiendo la habilidad que realmente nos proponíamos cuando diseñamos el test, o por el contrario los resultados de ese test nos están proporcionando información sobre otra habilidad distinta.

1.5. Principales tendencias y métodos de examen

Las principales tendencias que ha habido a lo largo de la historia de los tests usados en la evaluación de los idiomas son: la tendencia “precientífica”, la

tendencia “psicométrica-estructuralista”, y la tendencia “integradora-sociolingüística”.

Durante el periodo que duró la tendencia “precientífica” (hasta principios de los años 50), hubo poca preocupación por la fiabilidad, la validez o cualquier otra importante propiedad psicométrica de los tests. Las pruebas de examen estaban basadas en pruebas de lectura, traducción y gramática, pruebas que no habrían servido para medir de forma válida, la habilidad del candidato para usar la lengua y comunicarse en una situación de la vida real, fuera del contexto de la clase.

La segunda tendencia, la “psicométrica-estructuralista”, que duró hasta finales de los años 60, se basa en el enfoque analítico-estructural, que se estaba aplicando a la investigación lingüística del momento y que dio lugar a diversos estudios de análisis contrastivos. Estos estudios comparaban la fonología o la estructura del inglés con la del idioma del candidato y tanto la enseñanza como la evaluación se centraba en los puntos que creaban dificultad al alumno.

Para determinar de forma inequívoca el dominio o la falta del mismo de cada una de las áreas que habían sido identificadas a priori como problemáticas, se requirieron nuevos formatos de examen, ya que los métodos de preguntas abiertas de la etapa “precientífica” no servían. El contexto del test se redujo al mínimo y se crearon los exámenes llamados de elementos discretos (discret-point). El procedimiento que se siguió era “one-tested-element-per-test-item” Clark (1983: 432). Una de las características de los tests de elementos discretos es que intentan medir componentes específicos de una habilidad del idioma, tales como la gramática, el vocabulario o la pronunciación, de forma separada.

En 1914 y teniendo en cuenta lo anteriormente dicho, se introdujeron los tests de elección múltiple que permitía aplicarlos a gran escala y que fueron adoptados por la industria encargada de elaborarlos que surgió en los años 1920 y que pronto se convirtió en una gran empresa. Los tests de elección

múltiple fueron aceptados por la sociedad y por los tribunales de exámenes de distintas instituciones que los consideraban como herramientas objetivas y justas para todos los individuos, ya que proporcionaban técnica, eficiencia y estandarización a los exámenes. Al no ser criticados por la sociedad, estos tests se usaron como instrumento de poder que en un principio informaba a los políticos sobre las instituciones educativas pero que para finales de los años 70 se habían convertido en técnicas administrativas poderosas para controlar escuelas y sistemas educativos, a la vez que afectaba a la motivación, a los currículos y a la definición de conocimiento. Debido a su naturaleza, sin embargo, los tests de elementos discretos no eran capaces de medir de forma global la habilidad del alumno de comprender o producir.

La tercera tendencia, conocida como "integradora-sociolingüística", surge a finales de los 60. La investigación lingüística contemporánea sostiene que la comunicación en la vida real implica un acto creativo que no puede ser medido adecuadamente a través de la evaluación individual de las partes que lo componen. Los métodos de evaluación que se crearon basados en esta tendencia se llamaron pruebas integradoras. Entre estas pruebas se encontraban el dictado, *el cloze* y otras pruebas de redundancia reducida. Los contextos del examen deberían poseer genuina relevancia comunicativa para los candidatos, y los tests combinaban gramática y contexto, estructura y situación para proporcionar en la mayor medida posible situaciones que reflejaran la vida real.

En los últimos 20 ó 30 años las pruebas de idiomas han florecido proporcionándonos mejores herramientas para la administración, la corrección y el análisis de las mismas, dando respuesta a numerosas preguntas de investigación y creando otras nuevas.

En los años 70 la habilidad lingüística era vista como un conjunto de componentes finitos (vocabulario, gramática, pronunciación y ortografía) que se realizan en cuatro destrezas (comprensión lectora, comprensión auditiva, expresión oral y expresión escrita) que se evaluaban independientemente. La principal preocupación de los investigadores era la fiabilidad psicométrica de

las pruebas y se pensaba que la competencia del idioma consistía en un único rasgo unitario.

En los 80 el modelo de competencia comunicativa, que se dividía en cuatro componentes: competencia lingüística, competencia sociolingüística, competencia del discurso y competencia estratégica, fue muy influyente y forzó a las personas que preparaban las pruebas a abandonar la concepción de habilidad lingüística como un rasgo aislado, y a considerar "the discoursal and sociolinguistic aspects of language use, as well as the context in which it takes place" (Bachman, 2000: 3).

En los 90 nuevas metodologías tales como "item response theory", teoría de la generalización y de la medida referenciada a un criterio ("criterion-referenced measurement and generalizability theory) ocupó el lugar de la clásica fiabilidad referenciada a la norma (norm-referenced reliability). Bachman (1990), que estaba interesado en la relación entre competencia y lo que se hace, crea el modelo de competencia comunicativa en la cual las destrezas y los factores metodológicos conectan con lo que se hace en la vida real mientras que las competencias básicas están relacionadas con habilidades generales. Los enfoques cualitativos y cuantitativos de investigación se utilizan de forma complementaria y se incrementan las pruebas de idiomas con fines específicos.

En este siglo se espera un aumento en la variedad de los formatos de las pruebas y los procesos de evaluación disponibles debido al uso de sofisticados programas de ordenador, los cuales permitirán que proporciones considerables de datos sean manejados de forma rutinaria.

Bachman aboga por la formación de futuros profesionales de evaluación de idiomas y el uso de un código de práctica profesional. Más recientemente, el código de validez de Weir (2005a) considera lo que él llama "consequential validity" y Bachman llama "impact", que son las consecuencias y las implicaciones éticas que se derivan del uso de las pruebas de examen.

Bachman cree que la validez y la imparcialidad definirán el paradigma dominante para las pruebas de lengua en los próximos 20 años:

> Our increased sensitivity to ethical issues and the diversity and sophistication of our approaches to research mean that we can now research more than test scores, and we can go beyond speculation about their meaning and use. We are now in the position to assure ourselves and other test users that what we count counts. (Bachman, 2000: 25).

Todo este proceso de evolución y cambio en la evaluación de idiomas podemos verlo en la transformación que ha sufrido uno de los exámenes de inglés como lengua extranjera más clásicos de Gran Bretaña: "The Cambridge Proficiency Examination" (CPE). Este examen puede ser representativo del cambio en la enseñanza y evaluación del idioma inglés como lengua extranjera en Europa a lo largo del siglo XX. Cuando comenzó el CPE en 1913, el examen consistía en una prueba de literatura inglesa, una redacción, una prueba de fonética, una sección de gramática y traducciones directas e inversas en francés y alemán. El examen oral constaba de un dictado, una lectura en voz alta y una conversación. Esta división entre el examen oral y el examen escrito desapareció a partir de 1975 cuando se introdujeron 5 pruebas: redacción, comprensión lectora, uso del inglés, comprensión auditiva, y una entrevista.

Esta ruptura significó el reconocimiento de que la competencia de la lengua (language proficiency) no es unitaria sino parcialmente divisible. Hasta entonces, al menos en Gran Bretaña, se tenía la firme convicción de que la competencia de un idioma era unitaria y que, por lo tanto, importaba poco lo que se evaluaba siempre que se hiciera de forma fiable.

En los años 80 y 90 se logró una cierta convergencia internacional sobre la evaluación gracias *al Language Testing Research Colloquium*, que reunía anualmente a los profesores e investigadores interesados en la evaluación del idioma. En 1980 se creó también la revista *Language Testing* por un grupo de investigadores británicos de la Universidad de Lancaster, lo cual ha favorecido

el intercambio de puntos de vista entre ambos lados del Atlántico. En este momento los canales de comunicación están abiertos y se trabaja conjuntamente en el desarrollo, administración y análisis de los tests de idiomas.

1.6. Ventajas y desventajas de los tests

1. Los tests permiten fijar la nota de corte y que ésta sea distinta según el propósito y los resultados que se quieran obtener. Esto puede originar ideas distorsionadas de la realidad, ya que los tests o sus resultados pueden ser manipulados y de esta forma crear cuotas de forma flexible y arbitraria. En las Universidades de España tenemos como ejemplo las notas de corte exigidas para entrar a estudiar distintas carreras, como si hubiera alguna diferencia real entre los alumnos que tienen una nota de 6,9 ó los que tienen 7,1 cuando en una facultad se establece, por ejemplo, una nota mínima de 7,0 para poder estudiar tal o cual carrera. Otro ejemplo lo tenemos en los exámenes de oposiciones para acceder a un puesto de trabajo en la Administración Pública, en los que una diferencia de una centésima en la nota final puede hacer que una persona pueda conseguir o no una plaza.

2. Los temas y contenidos de los tests pueden llegar a ser más decisivos que el currículo del curso. Se puede llegar a controlar la enseñanza, el aprendizaje y el conocimiento de un idioma a través de los tests.
   Broadfoot, en la conferencia que dio en el Language Testing Research Colloquium en julio de 2003, alerta del peligro que existe de que los exámenes se conviertan en la finalidad última de la educación en lugar de intentar conseguir objetivos más amplios tales como, desarrollar y mejorar el conocimiento y las capacidades de los individuos para que puedan aprovechar las oportunidades de aprendizaje que se les presenten a lo largo de su vida.

> To improve learners´ knowledge and understanding, their skills and attributes.…to develop the capacities that will underpin learning throughout life so that individuals will be equipped to take advantage of the many new forms of learning opportunity that are now becoming available. (Broadfoot, 2005: 124)

3. Los tests simbolizan el orden social para el público en general, que siente un gran atractivo por ellos, especialmente los padres que los perciben como un signo de disciplina y educación de calidad. Este atractivo por los tests existe incluso entre las minorías, entre los emigrantes, entre personas con menor formación o entre las personas de clase trabajadora, aunque muy frecuentemente sean utilizados en su perjuicio por grupos políticos y empresariales.

1.7. Consecuencias de los exámenes a gran escala

Messick (1981, 1996) fue el primero en expresar que el impacto o las consecuencias de un test deben formar parte del concepto de validez del mismo. Los tests pueden tener consecuencias educativas, por ejemplo, cambios en los métodos de enseñanza, en las estrategias de aprendizaje, en los currículos, en los materiales usados en la enseñanza del idioma, en las prácticas de evaluación y en el conocimiento evaluado. Los exámenes también pueden tener consecuencias o efectos sociales en aspectos tales como la ética, la moral, la justicia, la ideología o las barreras sociales.

Existen al menos 4 términos para referirse a los cambios de comportamiento que pueden originar los tests: *impacto*, *consecuencias*, *efecto* y *washback* o *efecto rebote*. Aunque a veces se utilizan de manera indistinta, el *efecto rebote* se refiere a la influencia que los tests pueden tener sobre los contextos educativos, principalmente, enseñanza y aprendizaje.

Impacto es un término más general y abarca tanto el efecto que los tests puedan tener sobre la educación como sobre la sociedad. Messick dice que el *impacto* incluye la evaluación de las consecuencias de los tests, la imparcialidad en su uso y el efecto rebote que puedan tener en la enseñanza y en el aprendizaje

El término *consecuencias* engloba a los dos anteriormente descritos, centrándose especialmente además en los aspectos ideológicos. Las consecuencias de un test forman parte de la validez del constructo, creando lo que se ha llamado "consequential validity" o validez consecuencial.

A veces se introducen tests en la enseñanza y se utilizan como parte de las variables que contribuyen a la relación entre la enseñanza y el aprendizaje. El término que se utiliza para referirse a esta conexión entre la instrucción y el aprendizaje es conocido como "validez sistémica". Estos tests forman parte de una dinámica en la cual se realizan cambios en la enseñanza de acuerdo con la información obtenida por los propios tests, ayudando así a la mejora de la enseñanza del idioma. Es decir, que estos tests tienen también un claro impacto sobre el aprendizaje.

Uno de los efectos más comunes que producen los exámenes, especialmente los que tienen un nuevo formato para el alumno y que no ha podido practicar, es la ansiedad la cual va a afectar seriamente a sus resultados. En muchos países se ha introducido una cantidad ingente de tests para supuestamente mejorar la calidad de la enseñanza. Como consecuencia de ello muchos alumnos sufren de estrés y ansiedad, al tener que realizar un número de exámenes como el que no se había dado nunca en nuestros sistemas educativos. Los periódicos de Gran Bretaña vienen informando del aumento drástico de adolescentes que están tomando medicamentos, por ejemplo Prozac, como una ayuda para realizar los exámenes.

Alderson y Wall (1993) afirman que los buenos tests tienen obligatoriamente un efecto positivo y que los malos tienen un efecto negativo sobre la enseñanza, los profesores y los alumnos. Sin embargo, otros autores

están convencidos de que no hay ningún test que pueda ser considerado como bueno. Spolsky (1995: 56), por ejemplo, afirma que los tests siempre tienen un efecto negativo sobre la educación, ya que tienden inevitablemente a limitarla. Según él, no hay profesor que, ante un examen, no enseñe a sus alumnos el evitar cometer errores que puedan ser castigados severamente y a realizar prácticas que puedan ser premiadas. Así mismo se concentra en la clase de ejercicios que van a formar parte del examen, con frecuencia requerido también por los propios alumnos. Con esto, lo que está haciendo es enseñar a sus alumnos estrategias de examen más que un idioma.

Muy a menudo las autoridades educativas no cuentan con los profesores a la hora de introducir un nuevo examen. El papel de los profesores se limita a seguir órdenes, lo cual es muy frustrante al aumentar su responsabilidad y disminuir su autoridad. Además el examen se convierte en la fuente pedagógica más importante, empobreciendo el conocimiento, aún cuando las notas puedan ser más altas, ya que lo que ocurre es que se enseña para aprobar el examen y los alumnos estudian para sacar los mejores resultados posibles en el mismo y no para adquirir mayores conocimientos. Los tests se utilizan para establecer la posición de una escuela dentro del ranking de escuelas de una comunidad o de un país y también para clasificar el nivel educativo de un país con relación al de otros países. Este ritual parece servir para diferenciar los buenos de los malos, ya que cada vez que uno mejora de posición algún otro empeora. La función de estos tests sirve para fortalecer o menoscabar la reputación de profesores y escuelas.

Según Shohamy (2001a, 2001b), los tests no mejoran el aprendizaje ni nos ayudan a alcanzar un mayor rendimiento, y de hecho es una forma no democrática y poco ética de hacer política. A veces los tests que son útiles para las personas que toman las decisiones no lo son para los profesores o los alumnos.

Choi (2008) nos demuestra en su estudio las consecuencias tan negativas que los exámenes de EFL han tenido en la educación en Korea. A los estudiantes se les fuerza a emplear estrategias de examen cuando se les

prepara para entrar a la universidad. De acuerdo con su investigación el preparar a los estudiantes para realizar exámenes de elección múltiple en clase les priva de las oportunidades necesarias para aprender habilidades productivas. Estos alumnos no adquieren ninguna competencia comunicativa y son incapaces de aprobar el, para ellos, esperado examen de “English Proficiency”. Es por ello que la mayoría de los coreanos estén desilusionados con el examen de EFL. Sin embargo, los tests pueden ser muy útiles cuando están conectados con el aprendizaje que ocurre en las clases, ya que puede contribuir a mejorarlo, cuando la información que de ellos obtenemos se utiliza adecuadamente.

1.8. Usos de los tests

Los tests se utilizan para muy distintas finalidades tanto en las administraciones y empresas públicas como en las privadas. Algunos de los motivos por los que las Administraciones introducen un nuevo examen pueden ser:

1. Aumentar el prestigio de un idioma aumentando la motivación de profesores y alumnos.

2. Intentar cambiar los métodos de enseñanza sin apenas invertir económicamente en ello, ya que se evita tener que formar al profesorado. Muchos profesores reconocen que los exámenes modifican su forma de enseñar al prestar más atención a los temas y formato de examen. Esto puede ser positivo o negativo dependiendo del tipo de examen.

3. Fijar prioridades educativas.

4. Cambiar las estrategias de aprendizaje de los alumnos.

5. Aumentar el nivel de los alumnos.

6. Discriminar a ciertos ciudadanos por razón de idioma.

7. Aumentar la dificultad de acceder a ciertos trabajos, incrementando de este modo el prestigio social de esos trabajos.

8. Reducir la inmigración.

9. Sacar rendimientos políticos de forma más económica para los gobiernos, ya que los cambios de examen exigen menos inversión y son más visibles que formar al profesorado o cambiar los contenidos y currículos de un idioma.

10. Aumentar el número de personas matriculadas en un determinado curso.

Shohamy (2001a) nos da un ejemplo de cómo los tests se pueden utilizar como instrumentos de poder y control para cambiar el comportamiento de las personas de acuerdo con ciertas agendas que se desean introducir, y nos explica el proceso en la figura de la siguiente página.

(A) The origin of the power of tests

The detrimental force of tests along with their features of power cause those who are affected by the results of tests to change their behaviour and comply with the test's demands in order to maximize their scores and gain the benefits associated with high scores.

(B) Manipulations

Being aware of the capability of tests to affect behaviours leads those in power to introduce tests as means of creating and imposing changes in behaviours in line with specific agendas.

(C) Effects

Such use of tests has effects, yet the type and size of effects are complex and dependent on multiple factors such as status of the topic, purpose of the test, skill tested and whether the test is of high or low stakes,

(D) Consequences

The consequences of such uses of tests for education and society are a grater focus on the topic, narrowness of the knowledge, unethical behaviours, redefinition of knowledge, punishment, gate-keeping and controlling of education.

Figura 2. The power of tests – origins, manipulations, effects and consequences. (Shohamy, 2001a: 107)

Resumiendo, los exámenes son herramientas poderosas que juegan un papel muy influyente en nuestra sociedad y que pueden ser utilizadas para castigar, clasificar, disuadir, culpar, estandarizar, etc., sin que a veces se den cuenta ni las personas que los preparan ni las que se someten a ellos.

1.9. Evaluación crítica

Recientemente están apareciendo una serie de artículos y publicaciones, por ejemplo, Spolsky (1995) y Davies (1997) que critican los tests de idiomas y los colocan dentro del campo más amplio de la "Pedagogía Crítica". Es lo que se ha llamado "Evaluación Crítica" donde se debe tener en cuenta una serie de factores, tales como: qué se evalúa y para qué, qué no se evalúa y por qué no, quién evalúa a quién se evalúa, cuál es el efecto rebote, para qué se utilizan los resultados, cuáles son las consecuencias, que decisiones se van a tomar basándose en los tests, qué método se va a utilizar etc.

La Evaluación Crítica, por tanto, establece un diálogo social debatiendo sobre las prácticas y formas de evaluar el idioma así como la enseñanza y el aprendizaje del mismo. Además defiende la necesidad de aplicar procedimientos, de forma que la sociedad pueda protegerse del uso incorrecto de los tests. De acuerdo con Davies los enfoques críticos y alternativos para evaluar la lengua demuestran la importancia de poder demostrar que los tests que se usan son válidos.

> Critical approaches to language testing expose the importance of carefully examining alternative assessment proposals and making clear the validity of the assessment methods used by the profession. (Davies, 1997: 328)

Davies está a favor de establecer conductas de comportamiento explícitas que definan las obligaciones éticas y morales además de profesionales para todas las personas involucradas en la evaluación de un idioma, tales como profesores, autoridades, políticos, directores de centros, expertos en la elaboración de exámenes, correctores etc.

> Following critical theorists in other social sciences, critical applied linguistics have been asking questions about the ethics of applied linguistics and whether an ideologically neutral study of applied linguistics is possible. And where critical applied linguistics goes,

> critical language testing follows. There is urgent reason therefore both to examine the state of ethics in academic language testing and language research and to encourage a move towards explicit statement of good conduct practice. (Davies, 1997:329)

Los usos y consecuencias de los tests se están empezando a tomar en cuenta por algunas asociaciones tales como “The International Language Testing Association” (ILTA), y “The Australian Council of State School Organizations and the Australian Parents´Council” que han adoptado códigos de ética que sirvan como guía para los profesionales involucrados en la evaluación educativa. Cada vez son más frecuentes las publicaciones preocupadas por la evaluación crítica desde un punto de vista ético y de responsabilidad social. Sirvan como ejemplo: Hanson (1993), Hamp-Lyons (1997), Lynch (1997, 2001), Mcnamara (2001a, 2001b), Shohamy (2001a, 2001b), Choi (2008) y sobre todo el libro “Language testing: The social dimension” de McNamara y Roever (2006*)*. Por otra parte, las personas que trabajan en la educación se involucran cada vez más en la misma y toman medidas en contra de la imparcialidad de algunos tests. Incluso ya se está viendo que algunos de los casos de discriminación e injusticia están llegando a los tribunales (Lynch, 1997).

1.10. La ética en la evaluación de idiomas

Los dilemas morales han existido siempre en nuestra sociedad. Fulcher (1999b) nos dice que si la ética es simplemente un asunto de conveniencia social, el comportamiento sólo podrá ser regulado por convención. El problema es que la convención puede variar de sociedad en sociedad, y de comunidad en comunidad, con lo que la ética sería una cualidad relativa.

En los años 50 se originó un giro filosófico sobre el conocimiento. Como muchas otras cosas el conocimiento se transformó en un elemento de consumo y cada uno puede elegir el que desee. Hoy asumimos que no hay explicación

para todo, y para muchos Internet es la fuente principal de conocimiento. Lo mismo ha ocurrido con la ética y con el resto de los valores que también son relativos, temporales, locales y sin base lógica. Los problemas morales afectaron también a lingüistas e investigadores y el decimonoveno *Language Testing Research Colloquium* de 1997 se celebró con el título de: "Fairness in Language Testing". Desde entonces numerosos artículos han aparecido en la revista *Language Testing.*

Hamp-Lyons (1997) afirma que es comprensible que no existan principios éticos absolutos que nos lleven a comprometernos con una filosofía moral que considere la justicia o la imparcialidad ("fairness") en la evaluación. También se pregunta si las personas u organizaciones que elaboran los tests son responsables de su uso posterior.

- If tests are used to keep people out of countries, jobs or education, but this was not their stated purpose when designed, is it the responsibility of the test developers?

- Are testing organisations (ETS/UCLES) responsible for the use of test scores to make decisions that they say they should not be used to make?

- Are testing organizations responsible for any damage caused by cramming schools that earn large amounts of money out of providing test practice, rather than teaching?

Ella no contesta a estas preguntas pero deja entrever que la ética es diferente en diferentes sociedades y que puede cambiar según las circunstancias.

Davies (1997) no opina lo mismo que Hamp-Lyons (1997) y cree que el ser profesional está relacionado con los códigos, los contratos, la experiencia y los niveles de práctica. Estos pueden cambiar a lo largo del tiempo pero la profesión debe comportarse de acuerdo con unas normas independientemente del lugar donde se viva, de la nacionalidad y para que institución se trabaje.

Las personas que elaboran test de idiomas han venido aceptando las normas éticas reflejadas en varios documentos. Entre los más conocidos están: “The Code of Fair Practices in Education”, elaborado por The National Council on Measurement in Education, o el código de práctica de la “International Language Testing Association” (ILTA). Además, consideran que el APA: Standards for Educational and Psycological Testing (1985) contiene las reglas básicas para sus actividades.

Otro tema que ha atraído mucho la atención de los investigadores es la responsabilidad de los que elaboran un test, sobre todo, los que participan en proceso de evaluación. Esto incluye a las personas que se examinan, a los profesores, a las escuelas que administran los exámenes, a los oficiales públicos etc. Se trata de intentar asegurar un impacto positivo del test sobre todos ellos.

Finalmente, el efecto rebote o “washback” también hay que tenerlo en cuenta debido a las consecuencias éticas que el test puede tener en la práctica docente. Este efecto fue formulado por Wall y Alderson (1993) e indicaba que la relación entre la evaluación y la enseñanza es más compleja de lo que se había pensado. Desde entonces el efecto rebote ha seguido siendo investigado: Alderson y Hamp-Lyons (1996); Hamp-Lyons (1997) o Amengual (2009). Algunos autores han ampliado la noción de efecto rebote y lo han llamado “impacto” (Wall, 1996; Bachman y Palmer, 1996; McNamara, 1996, 1998; Shohamy, 2001a, 2001b y Weir, 2005a), el cual busca investigar la relación existente entre el uso de un test y la sociedad en la que se usa.

Todo este debate ha ampliado el concepto de evaluación del idioma, que incluye no solamente aspectos técnicos de desarrollo y administración de un test sino que hay que tener en cuenta el contexto en el que el test se va a realizar y administrar. Para ampliar el entendimiento del uso de los tests de idiomas en una gran variedad de contextos se está echando mano de la filosofía, de la ética, de la teoría crítica y de los estudios políticos sociales. Estos temas ya fueron tratados por Messick (1981), quién incorporó la ética en el concepto de validez del test y más específicamente en la validez del

constructo. Después fue seguido por otros autores tales como Shohamy (2001a, 2001b) o Weir (2005a).

De acuerdo con Fulcher tenemos que esforzarnos en que los tests sean justos e imparciales y desarrollar un criterio para decidir cuando y como es correcto el uso de los tests:

> The whole history of the study of reliability is essentially a striving for "fairness". It is a great pity that many of our largest examination boards still do not understand the concept, let alone calculate it. …… We need to plot the impact that tests have on tests takers and the societies in which they are used. And we need to develop criteria to decide when and how test use is right or wrong.
> (Fulcher, 1999b)

1.11. El futuro de los tests

Existen diferentes puntos de vista sobre el futuro de los tests. Algunos creen que la era de los tests se ha acabado, ya que son herramientas demasiado autoritarias para que puedan ser utilizadas en sociedades multiculturales que demandan legitimidad, respeto, identidad y derechos. Un ejemplo de esta corriente es Valdés y Figueroa (1996). Ellos defienden que los tests son instrumentos de estandarización de poblaciones enteras, los cuales proporcionan información que no es fácilmente interpretable y sin embargo, tienen muchas implicaciones para muchos individuos, así mismo, están de acuerdo con la opinión publica de que los tests no pueden medir el conocimiento de forma precisa.

Otra corriente, entre los que se encuentra Broadfoot (1996), Creen que los tests deberían continuar existiendo y ejercer más control que nunca. Opina que los tests pueden ser útiles para oponerse a los que exigen compartir el poder y que es la combinación de tecnología y burocracia lo que permitirá a

ciertos grupos aumentar su poder gracias a que sus decisiones políticas estarán legitimadas al estar basadas en un proceso objetivo y racional. De esta forma, el individuo no podrá oponerse a los efectos cada vez más intrusos de la maquinaria del estado.

Y por último están los que opinan que los tests se quedarán con nosotros pero de forma diferente. Manifiestan que los tests pueden ser beneficiosos y constructivos, y gracias a la información que de ellos recibimos podemos enseñar y aprender de forma más eficaz. También se pueden utilizar como protección en contra de formas de poder y control poco éticas o democráticas. Madaus (1990) no se imagina que los tests puedan ser abolidos en una sociedad donde se utilizan para promover y rechazar ciertos valores específicos y relaciones sociales. Sin embargo, recuerda que los tests, lo mismo que el resto de las tecnologías importantes, deben ser sometidos a una evaluación.

Shohamy es de la opinión de que los tests que utilicemos deben ser de calidad y cumplir todos los criterios de validez por mucho que cueste.

> There is a need for continuous examination of the quality of tests, for the in-depth insight into how they are used, for the public exposure of misuses and for the awareness of the public as to the motivations for, harm and consequences of tests. The cost of this approach is high – it takes more time, it involves more people, it requires greater resources. It requires compromise as all democratic practices do. But, if tests are so central, yet they pose such strong potential for misuses, the cost is worth paying.
> (Shohamy, 2001a: 161)

Así pues, los tests del futuro, que administrados responsablemente, van a seguir siendo una fuente importante de información, además de fiables y válidos deberán ser utilizados de forma educativa, ética y democrática, es decir, involucrando y colaborando con el mayor número posible de personas y estamentos. Ahora tenemos evidencia de que los tests no son herramientas

neutrales por lo que cuando confeccionamos un test deberemos prestar atención a las intenciones que existen al administrar ese test, sus efectos y sus consecuencias y considerar estas variables al definir la validez de los tests. Temas éticos importantes, junto con otros temas de validez, se articularán desde una perspectiva crítica y se separarán del enfoque más tradicional de la evaluación de idiomas (Lynch, 2001).

Un área importante de investigación puede ser la influencia del método sobre los resultados de los tests. Es importante que los tests de idiomas maximicen la influencia de la competencia que quieren medir y minimicen la forma de medirla. Por ejemplo maximizar la competencia de lectura en un examen de comprensión lectora, y minimizar la influencia de, digamos, la capacidad de hacer tests de elección múltiple. (Skehan, 1991).

Otra deseable área de progreso, que en parte ya ha empezado desde que disponemos de tests por ordenador, es el desarrollo de los tests de autoevaluación. La filosofía de la enseñanza y del aprendizaje de idiomas está evolucionando. Ahora es una actividad que se desarrolla a lo largo de toda nuestra vida por lo que es de suma importancia el conseguir que las personas que deciden aprender un idioma puedan hacerlo de forma autónoma. El reto es mantener la relevancia y utilidad de los tests de idiomas en medio del profundo cambio intelectual y político que se está produciendo.

1.11.1. Los tests por ordenador

No podemos hablar del futuro de los tests sin hablar de las nuevas tecnologías. Con la expansión de la enseñanza a distancia, el desarrollo de internet, con la mayor disponibilidad y accesibilidad tanto al software como al hardware, los ordenadores son ahora ampliamente usados en la educación. En 1998 se introdujo una versión para ordenador del Test of English as a Foreing Language (TOEFL). Sin embargo, todavía no está claro hasta que punto son

solamente ayudas para presentar, administrar y analizar los tests o representan cambios en el método y contenido. (Davies, 2003).

Los tests por ordenador tales como los conocidos, Computer-Based Language Testing (CBLT), Computer-Assisted Language Learning (CALL); (Alderson y Windeatt, 1991) o Computer-based tests (CBT) tienen las siguientes desventajas:

1. Hay que tener experiencia en el manejo del ordenador para no estar en desventaja.

2. No permite mucha variedad de métodos de examen. Generalmente se abusa de la técnica de elección múltiple así como de los métodos de rellenar huecos o del *cloze*. Ello es debido a que las respuestas tienen que poder ser corregidas a máquina.

3. Las destrezas productivas de expresión oral o expresión escrita no pueden ser evaluadas de forma significativa.

4. Existe el problema de la seguridad de la base de datos. Por ello todavía se utilizan principalmente para exámenes que no son a gran escala.

5. No se pueden cambiar las respuestas aunque se cambie de idea.

Sin embargo, los tests por ordenador también ofrecen ventajas significativas, con lo que se augura que en el futuro aumentarán considerablemente. Entre las principales ventajas están:

1. No es necesario fijar un día o lugar concreto para el examen. Existen exámenes disponibles a cualquier hora del día o de la noche.

2. Los resultados pueden estar disponibles inmediatamente después del examen.

3. La base de preguntas puede ser actualizada constantemente por expertos que tienen acceso a internet independientemente de su situación geográfica.

4. Se pueden utilizar para autoevaluación y para exámenes de diagnóstico.

5. Puede ser un método factible e inteligente de acceder a exámenes justos para un gran número de personas.

Uno de los exámenes que están disponibles en internet es DIALANG (Alderson y Huhta, 2005). Éste es un examen de diagnóstico que contiene tests de comprensión lectora, comprensión auditiva, expresión escrita, vocabulario y estructuras en 14 idiomas europeos, y es gratis. El marco y las especificaciones del examen correspondieron al Marco Común de Referencia de la Unión Europea, que fue el que dio los fondos para esta iniciativa, juntamente con instituciones y universidades de toda la Unión Europea. Los resultados se dan en los 6 niveles de la escala del Marco Común de Referencia, que van desde el A1 hasta el C2. DIALANG es el primer examen de idioma a gran escala cuyo objetivo es diagnosticar, no certificar la competencia de un idioma. A pesar de las ventajas de los tests por ordenador Davies (2003) es escéptico y señala que podemos estar ante un caso en el que permitamos que la herramienta domine el contenido y el constructo.

Finalmente, diremos que la reciente investigación sobre la evaluación de idiomas ha intentado descubrir los procesos y estrategias que utilizan los candidatos cuando responden a las tareas de los tests, y el único resultado que se ha obtenido es que la variedad de aproximación de los individuos a los tests es inmensa, lo que significa que los mismos aciertos pueden representar diferentes habilidades, o diferentes combinaciones de habilidades, y resulta actualmente imposible definir exactamente lo que significa una nota. Es lo que Alderson y Banerjee (2002) llaman “The Black Hole of language testing”. Según ellos todavía existen dilemas a la hora de medir la competencia en un idioma y lo emocionante acerca del estado actual de la evaluación de un idioma es el

darse cuenta de lo poco que sabemos sobre el efecto rebote, la innovación, el comportamiento ético o lo que estamos midiendo con un test.

> We know less than we think, whether it is washback, politics and innovation, ethical behaviour, what exactly we are testing, or how to know what we are testing. The challenge for the next decade will be to enhance our understanding of these issues.
> (Alderson y Banerjee, 2002)

1.12. Clases y uso de las pruebas de idiomas

La clase de pruebas que hacemos dependerá mucho de nuestros propósitos de evaluación. Como ya se ha dicho anteriormente, el principal propósito de los tests es proporcionar información para la toma de decisiones. Tanto en nuestra sociedad en general como en los programas educativos en particular, nos interesan principalmente dos clases diferentes de decisiones: sobre individuos y sobre programas, que Bachman (1990: 58) llamaba micro-evaluación y macro-evaluación respectivamente.

De acuerdo con nuestros fines tenemos las siguientes clases de pruebas:

1.12.1. Pruebas de clasificación

Nos permiten agrupar a nuestros alumnos homogéneamente. Por lo tanto se concentran en evaluar un amplio y representativo rango de destrezas generales de la lengua. Generalmente son pruebas hechas a medida, de forma que podamos obtener un amplio abanico de notas que nos ayude a dividir a los alumnos en varios grupos de acuerdo con su nivel de habilidad de la lengua.

1.12.2. Pruebas de diagnóstico

Se utilizan para intentar identificar áreas específicas que necesitan ser reforzadas o explicadas de nuevo. Generalmente están basadas en el currículo y nos ayudan a identificar los puntos fuertes y débiles de nuestros alumnos y cuales son sus dificultades de aprendizaje. Muy a menudo las pruebas de diagnóstico se realizan a través de *Pruebas de Auto-evaluación,* especialmente ahora que disponemos de pruebas de ordenador de libre acceso y con información inmediata como son las de DIALANG (Alderson y Huhta, 2005) que se han comentado anteriormente.

1.12.3. Pruebas de admisión

Estas pruebas se utilizan en muchos países como un complemento de otras clases de información. Un ejemplo es los exámenes de entrada a la Universidad, en España llamados Pruebas de Selectividad (Herrera 1999, 2005; Herrera et al. 2001), en las que los exámenes ayudan a decidir que alumnos son aceptados en ciertas carreras o en ciertos programas académicos.

1.12.4. Pruebas de progreso o pruebas formativas

Las pruebas formativas son las realizadas en clase por los profesores para promover el aprendizaje. Nos permiten conocer hasta que punto los alumnos dominan los contenidos y destrezas que se les ha enseñado anteriormente. Así mismo informan a los profesores sobre la conveniencia de las actividades que se hacen en clase. La información recibida debería servir para activar procesos adecuados de enseñanza y aprendizaje. Heaton (1990) y

especialmente Rea-Dickins (2001) identifican los exámenes en clase como un fenómeno de múltiples facetas ligadas al aprendizaje, a la enseñanza, al cumplimiento del currículo y a las funciones burocráticas. Afirman que estas pruebas se pueden utilizar para mostrar a los alumnos el progreso que han conseguido y así aumentar su motivación, para involucrar a los alumnos en el aprendizaje del idioma, y para aumentar su confianza y su autoestima.

1.12.5. Pruebas de rendimiento

Son como las de progreso pero abarca un periodo de aprendizaje más amplio. Generalmente son exámenes formales realizados al final del curso para determinar si los alumnos dominan o no el contenido del mismo. Las pruebas de rendimiento están basadas bien sobre el currículo y los contenidos del curso, o bien sobre el libro de texto que se ha usado. (Hughes, 1989: 11) se refiere a esto como “syllabus-content approach”. Davies (1985) afirma que la mayoría de los tests en la enseñanza son pruebas de rendimiento basados en un programa dado y un tipo de enseñanza determinado.

1.12.6. Pruebas de selección

Estas pruebas están diseñadas para seleccionar ciertos candidatos para un trabajo o para una plaza en un curso. Los resultados se interpretan de acuerdo con el comportamiento de todo el grupo, de ahí que en inglés se llamen “norm-referenced tests” y en español pruebas normativas o pruebas referenciadas a la norma. Los profesores, los médicos y el resto de los funcionarios locales, autonómicos y estatales en España, por ejemplo, son seleccionados usando estas pruebas, ya que hay siempre mayor número de candidatos que puestos de trabajo o plazas ofertadas.

La finalidad de estos exámenes es, por lo tanto, comparar los resultados de todos los candidatos y seleccionar solamente a los mejores. El número de candidatos que se considera que han aprobado el examen es arbitrario y está determinado por la disponibilidad de plazas en una institución.

1.12.7. Pruebas de nivel

Se usan principalmente para descubrir lo que el alumno ha aprendido de un nivel de habilidad claramente definido o del contenido que ha sido impartido generalmente a lo largo de un curso. Tanto el nivel como el criterio de evaluación se definen a priori y cualquier candidato que alcance ese nivel se considera que ha aprobado. Sus notas se interpretan como indicadores de un nivel de habilidad o el grado de maestría del contenido del curso que han alcanzado.

1.12.8. Pruebas de competencia o de certificado

Son exámenes basados en la teoría y están diseñados para medir la habilidad de las personas en un idioma independientemente de los cursos o los años de estudio que hayan dedicado a ese idioma. Se utilizan para saber si los candidatos tienen suficiente dominio de la lengua para un propósito en particular de la vida real. Se basan en especificaciones de lo que los candidatos tienen que saber hacer con el idioma para que se les considere competentes en el mismo. Si las pruebas de competencia se basan en una teoría que no difiere de la que se basa un currículo entonces, las pruebas de competencia y las de rendimiento pueden ser muy similares.

1.12.9. Pruebas para fines específicos

Las Pruebas para Fines Específicos o ESP (English for Especific Purposes), pueden ser de admisión a unos cursos especializados, de rendimiento en esos cursos, de selección para un trabajo, y sobre todo de competencia. Son pruebas basadas generalmente en el análisis del lenguaje que se necesita para un propósito determinado. Son exámenes dirigidos a profesionales que tienen que demostrar que son capaces de usar la lengua en áreas específicas, como el turismo o los negocios. Si por ejemplo vamos a diseñar un test de competencia para controladores aéreos, tendremos que basarlo en las destrezas que se necesitan en la torre de control.

Dentro de los exámenes de inglés para fines específicos se encuentran también los exámenes de inglés para fines académicos o EAP (English for Academic Purposes). Como en todos los exámenes, antes de diseñarlo tenemos que describir a los participantes, analizar sus necesidades comunicativas, y especificar el contenido del test.

En los exámenes de ESP se tiene especialmente en cuenta el futuro trabajo o estudio de los candidatos. Es lo que se ha llamado como el enfoque de “la vida real”, en el que no solamente es importante el contenido del test sino también el formato del mismo. El contenido del test y el tipo de tarea refleja el uso de la lengua en situaciones en las que el candidato tiene que sobrevivir.

Las cualidades más importantes de estos exámenes son la autenticidad y la validez aparente. Para Bachman (1990: 307) “face validity is the appearance of real life and content relevance is the representation of real life and predictive utility is essentially precluded without authenticity”. La autenticidad en los exámenes significa por lo tanto, el grado en el que el mundo real se ve reflejado en las condiciones del test.

En cuanto a la influencia que el tema pueda tener sobre los resultados del test no existe evidencia clara de que la especificidad de los temas de los textos utilizados en un examen contribuya en gran medida a la nota final. Lo realmente decisivo en los resultados finales es el nivel de competencia que un candidato tenga del idioma. Fulcher afirma que en niveles intermedios el conocimiento del tema puede compensar la falta de nivel de un candidato lo que no ocurre cuando la competencia del alumno es más elevada.

> - Language proficiency accounts for most of the variance in EAP test scores.
>
> - Subject knowledge can compensate to a small degree for lack of language proficiency, but this appears only to happen at intermediate levels.
>
> - What makes a text specific to a subject area cannot be defined by expert judges.
>
> - Increased specificity by module proliferation is unnecessary.
>
> (Fulcher, 1999a)

Por lo tanto, las únicas razones para seguir manteniendo este tipo de exámenes específicos serían por una parte, la necesidad que tienen los candidatos de percibir el test como algo relevante para sus estudios o su trabajo para así conseguir una respuesta válida y por otra, el efecto rebote que el contenido, los títulos y el formato de los textos del examen pueden tener sobre lo que hacen los profesores en las clases.

Tanto Fulcher (1999a) como Clapham (2000) concluyen que los exámenes de ESP y especialmente los EAP, aunque pueden seguir siendo parecidos a los que hemos conocido hasta ahora, deberían ser calificados únicamente teniendo en cuenta habilidades generales del idioma. También sugieren que deberíamos llamarlos “tests of English through Academic Context” (EAC) en lugar de “tests of English for Academic Purposes” (EAP). Clapham

además añade que los exámenes deberían incluir una variedad más amplia de tareas que no tuvieran en cuenta exclusivamente el contenido. Entre sus sugerencias se encuentra un test de gramática que sería adicional al resto de tests comunicativos. Esto lo justifica principalmente por el efecto rebote del examen.

> If students are to function efficiently in an English-speaking academic speech community, they have to be able to recognise and accurately use a wide range of English structures. So some emphasis on grammatical skills might lead to *positive* washback. (Clapham, 2000: 518)

1.12.10. Pruebas directas

Las pruebas directas requieren que el candidato realice precisamente la destreza que deseamos medir. Se utilizan generalmente para medir destrezas productivas orales o escritas usando tareas y textos lo más auténticos posible. En las pruebas de competencia directa, el formato de examen y el procedimiento intentan imitar al máximo una situación de la vida real en la cual se demuestra normalmente la competencia. (Clark, 1975: 10; en Bachman, 1990: 304).

Las pruebas directas son más fáciles de diseñar y hoy en día son más aceptadas por los alumnos, ya que la tarea que tienen que realizar se parece más a las actividades de la vida real, por ejemplo, escribir una carta, contestar al teléfono, etc. Si quisiéramos evaluar la pronunciación de un alumno a través de una prueba directa, le mandaríamos hablar.

1.12.11. Pruebas indirectas

Las pruebas indirectas intentan medir las habilidades que subyacen en las tareas que proponemos a los alumnos. Las medidas indirectas no necesitan reflejar contextos en los que el idioma se use de forma auténtica, y en muchos casos, pueden tener muy poco parecido formal con las situaciones lingüísticas que el alumno va a encontrar en la vida real. (Clark, 1978b: 26; en Bachman, 1990: 304). Bachman nos recuerda que la relación entre las puntuaciones del examen y las habilidades que queremos medir es siempre indirecta.

> As with all mental measures, language tests are *indirect* indicators of the underlying traits in which we are interested, whether they require recognition of the correct alternative in a multiple-choice format, or the writing of an essay. (Bachman, 1990: 33).

Un ejemplo de prueba indirecta sería evaluar la pronunciación de un alumno con un test de elección múltiple, en el que el alumno tiene que identificar las palabras que tienen la misma pronunciación vocálica. El *cloze* test y el C-test constituyen también ejemplos de pruebas indirectas.

1.12.12. Pruebas de elementos discretos y pruebas integradoras o comunicativas

Cuando se tiene claro el aspecto de la lengua que se quiere evaluar y si se desea evaluar ese aspecto específico de forma separada, se utilizan pruebas de elementos discretos. Por ejemplo si deseamos evaluar exclusivamente ciertas estructuras gramaticales. Este método tiene la desventaja de que el efecto rebote no es muy positivo, ya que las pruebas no reflejan el uso real de la lengua. A pesar de ello estos tipos de pruebas, como por ejemplo los tests de elección múltiple, se siguen utilizando con mucha frecuencia debido a que son fáciles de preparar, administrar y corregir. El

número de preguntas puede ser elevado con lo cual, el coeficiente de fiabilidad suele ser alto.

En contraposición con las pruebas de elementos discretos, las pruebas integradoras, llamadas también comunicativas, nos ofrecen una visión global de la habilidad de un candidato. Para realizar estas pruebas el alumno tiene que hacer uso de varios elementos lingüísticos. Las pruebas integradoras correlacionan muy bien con otras pruebas y son aceptadas por los candidatos fácilmente, ya que reflejan situaciones reales de comunicación, cosa que no hacían las pruebas de elementos discretos. El problema que pueden presentar es el de la fiabilidad en la corrección, pero esto puede subsanarse como veremos más adelante. La mayoría de las pruebas de elementos discretos son pruebas indirectas, mientras que la mayoría de las pruebas integradoras son pruebas directas. Hay algunas excepciones tales como el C-test y el *cloze* que son pruebas indirectas e integradoras.

1.12.13. Pruebas subjetivas y objetivas

Dependen enteramente del procedimiento de corrección. El examen y la corrección son objetivos cuando el corrector no necesita emitir ningún juicio sobre si la respuesta es correcta o no. Por otra parte, si el corrector debe juzgar las respuestas, nos encontramos ante exámenes y correcciones llamados subjetivos. Los evaluadores buscan la objetividad debido a la mayor fiabilidad que ella aporta.

1.12.14. Pruebas de idioma comunicativas

Las pruebas de idioma comunicativas miden la habilidad de un candidato para formar parte en actos de comunicación incluyendo la comprensión lectora

y la comprensión auditiva. Se las describe como pruebas de competencia basadas en tareas, realistas, que usan materiales auténticos e integran no sólo destrezas sino también lengua. Se trata de evaluar si el candidato es capaz de usar la lengua adecuada para realizar ciertas tareas similares a las de la vida real. Porter (1991: 33) sugiere que las pruebas de lenguaje comunicativo deben estar basadas en las necesidades de las personas que se examinan y que tanto el contexto como el objetivo de la tarea deben estar adecuadamente incorporados en las pruebas.

1.13. Factores que afectan a la actuación de los candidatos en las pruebas de idiomas

Ya que la habilidad que un individuo tiene sobre una lengua no puede ser medida directamente, es preciso recurrir a los exámenes de idiomas para basándonos en los resultados de los mismos poder deducirla.

Para Bachman (1990) la *habilidad del lenguaje comunicativo* incluye *la competencia del idioma*. Sin embargo, para Bachman y Palmer (1996) la habilidad del lenguaje consiste en *el conocimiento del lenguaje* y en *la competencia estratégica.*

La actuación en los exámenes de idioma se ve afectada por una variedad de factores: condición física del examinando, alerta mental, edad, conocimientos propios, lengua materna, formato de examen, aptitud, interpretación de los resultados de la prueba, materiales utilizados en el examen, temperatura de la habitación donde se administra, si es una prueba directa o indirecta, objetiva o subjetiva, etc. Algunos de estos factores pueden ser controlados por los diseñadores y administradores del examen, pero algunos otros, tales como el género, los conocimientos culturales, las características personales, etc., sólo se pueden controlar en muy pocos contextos.

Nuestra mayor preocupación cuando diseñemos un examen, por lo tanto, es minimizar los efectos del método, de las características personales del individuo que no formen parte de la habilidad del idioma, y de otros factores aleatorios que puedan influir en la actuación de los candidatos. Así mismo, la corrección de las pruebas, la interpretación y el uso de las calificaciones deben ser analizados de modo apropiado, de forma que no reflejen otros factores que no sean la habilidad del idioma que queremos medir.

1.13.1. Competencia estratégica

La competencia estratégica es un factor no lingüístico en el uso de un segundo idioma. Compensa la insuficiencia de la competencia lingüística del lenguaje de los usuarios del idioma. Incluye estrategias de comunicación verbales y no verbales que están disponibles para el individuo que puede usarlas para compensar su falta de competencia o para resaltar su actuación. Hay algunos tipos de tareas que pueden ser completadas con éxito por los examinandos usando una competencia estratégica no verbal, en lugar de la habilidad específica del lenguaje que la prueba originalmente pretendía medir.

De acuerdo con Phakiti (2008), la relación entre la actuación de un candidato y su competencia estratégica es muy limitada, ya que el aprendizaje o uso de un idioma depende de muchos factores, por ejemplo: la dificultad de la tarea, factores cognitivos y metacognitivos, la competencia en el idioma, la memoria de los candidatos, los niveles de corrección con que un idioma se procese automáticamente o la motivación para usar el idioma. Por lo tanto, el hecho de poseer una gran competencia estratégica no conlleva el tener éxito en el aprendizaje y uso de un idioma.

1.13.2. Conocimiento del tema

Llamado también "conocimiento de los esquemas" o conocimiento del mundo real por Bachman y Palmer (1996: 65), permite a los examinandos usar el lenguaje relacionándolo con el mundo en el cual viven. Alderson (2000: 34) piensa que este concepto se puede dividir en "*background knowledge*" – que puede o no guardar relación con el contenido y el tema de la prueba y "*subject-matter knowledge"*, que está directamente relacionado con el contenido y el tema de la prueba. El conocimiento general o del mundo es esencial y tiene un efecto muy fuerte sobre los resultados del examen. Se ha demostrado claramente que el hecho de que un tema sea familiar para un individuo, mejora su actuación en las pruebas de idiomas.

El conocimiento del mundo en general se refiere a "tu mundo", es decir, a tu cultura. Por lo cual el conocimiento cultural es también crucial para entender la información. Sin embargo, se ha demostrado que los estudiantes maduros de un segundo idioma son capaces de compensar su falta de conocimiento cultural relevante o las diferencias en la forma en la que se organizan los textos en los dos idiomas, evitando así que su actuación sea obstaculizada por lo que los investigadores llaman esquemas de contenido o esquemas formales respectivamente. (Read 200: 192).

De acuerdo con Tarone (1998: 77), "even when tasks are held constant, topic can affect interlanguage performance". Esto se demostró con un grupo internacional de ayudantes de profesores que hicieron un examen oral y cuya actuación cambiaba drásticamente según se les diera un tema general o un tema específico para, por otra parte, tareas idénticas.

Alderson y Urquhart, en su investigación: "The Effect of Student Background Disciplines on Comprehension: a Pilot Study", analizan los resultados de un test de comprensión lectora que realizaron estudiantes universitarios de habla no inglesa en el Reino Unido. Querían demostrar que los estudiantes que realizaban exámenes de comprensión lectora basados en

textos que contenían un tema familiar, o sea, relacionados con su área de conocimiento, obtenían mejores resultados que los estudiantes que no conocían el tema del texto del examen. Eligieron cuatro grupos de estudiantes de diferentes áreas: desarrollo administrativo, ingeniería, matemáticas y físicas y ciencias sociales. La prueba consistía en varios *cloze tests* sobre textos de cada una de las cuatro disciplinas que estaban estudiando los estudiantes. Los resultados indicaron que, efectivamente, los candidatos obtenían mejores resultados cuando los tests se basaban en textos correspondientes a sus propias disciplinas:

> The hypothesis was supported that students from a particular discipline would perform better in tests based on texts taken from their own subject discipline than would students from other disciplines. That is, students appear to be advantaged by taking a test on a text in a familiar content area.
> (Alderson y Urquhart, 1983)

Sin embargo, Jennings et al. (1999) llevaron a cabo una investigación sobre el efecto del tema en la actuación de los examinandos usando el mecanismo de elección. Sus resultados sugieren que aunque los examinandos, con frecuencia sienten que el tema de los exámenes tiene un impacto sustancial sobre su comportamiento, los resultados no fueron significativamente diferentes. Sin embargo, no debemos olvidar los efectos emocionales y psicológicos que tiene el poder elegir el tema.

1.13.3. Esquemas afectivos

Determinan la respuesta afectiva del examinando hacia la tarea, la cual "may influence the ways in which they process and attempt to complete the test tasks". (Bachman y Palmer 1996: 66). Los evaluadores deberían ser conscientes del efecto que los temas cargados emocionalmente, tales como el aborto o la pena de muerte, tienen en las pruebas de idioma y el hecho de que

los examinandos puedan verse afectados. Tenemos que considerar eliminar de los exámenes todos los temas que puedan causar cualquier angustia a los examinandos, ya que es una fuente potencial de falta de fiabilidad. Porter (1991: 36) y Tarone (1998: 76) muestran que la cualidad de la actuación del lenguaje hablado puede variar de manera previsible con las características del *interlocutor,* punto fundamental que debe ser tenido en cuenta por los evaluadores.

1.13.4. Efecto del método

Está demostrado que la forma de la prueba tiene considerable influencia sobre los resultados. El comportamiento en los exámenes de idiomas varía dependiendo tanto de la habilidad individual en el idioma como de las características de la prueba o método empleado.

Bachman (1990:156-157) cree que la función de evaluación de la competencia estratégica proporciona un mecanismo para explicar esta variabilidad. Él describe un marco para caracterizar las facetas del método de examen que afectan al comportamiento en la evaluación del idioma y considera "the testing environment, the rubric, the input, the expected response, and the relationship between input and expected response", como las características de la tarea que afectan a los resultados de los examinandos. Muchos investigadores están preocupados por el efecto que puede tener el método sobre los resultados del test y la forma de eliminar ese efecto.

> The format of the test may itself intrude and cloud the measurement that is being made. This aspect of the model implies a recognition of the fallibility of testing, of the way in which part of a test result may be the result of test format effects rather than underlying ability, and most ambitiously, that testers need to know about systematic effects of these sorts if they are to allow for them

> in actual test results, or, better still, to avoid them. (Skehan, 1991: 10)

Murphy (1985: 15) considera que aunque hay un gran número de métodos cuantitativos y cualitativos disponibles, los evaluadores deben asegurarse de que todos ellos son fiables, válidos y accesibles, es decir, que se pueden describir y repetir para evitar distorsión. Bachman (1990) sugiere que una forma de escapar del dilema de evaluar un idioma – lo cual implica que el idioma es a la vez instrumento y medida – es entender mas explícitamente la naturaleza de la habilidad del lenguaje y el lenguaje de los métodos de evaluación para "minimizar el efecto del método" en la interpretación de los resultados como indicadores de las habilidades del lenguaje. Sin embargo, Douglas (1998: 153) ve la necesidad de "capitalizar" los efectos del método diseñando pruebas para poblaciones específicas.

1.13.5. Dificultad de las pruebas

Hay pruebas que son más exigentes que otras por las características de los textos o los ítems. Mantendremos el término inglés "*input*" para referirnos a la información que contiene una tarea determinada de un test a la que se enfrenta un candidato. La respuesta o la tarea será más o menos difícil dependiendo de las características del *input* como pueden ser el tema, la especificidad de la información, el nivel de vocabulario, la complejidad sintáctica, la concreción, las expresiones negativas o la cantidad de procesamiento requerida (Bachman, 1990).

1.13.5.1. La carga de vocabulario

La carga de vocabulario es el factor más importante a la hora de predecir la dificultad de un texto y puede variar de acuerdo a varios factores:

a) *El perfil de frecuencia léxica* que está basado en la frecuencia relativa de las palabras en el idioma. Las palabras polisilábicas – palabras de tres o más sílabas – son menos frecuentes que las de una o dos sílabas y cuanto menos frecuente es el vocabulario del texto que proporcionamos al alumno más difícil será la tarea que tienen que realizar.

b) *La densidad léxica* es la proporción de palabras de contenido léxico, lo que en inglés se llama "lexical or content words", en el texto. Las *palabras de contenido léxico* son las palabras cuyo significado mejor describe el diccionario y pertenecen a la clase abierta. Esto quiere decir que nuevas palabras, tales como palabras técnicas, palabras adoptadas o adaptadas de otros idiomas, palabras que provienen de distintas jergas - jerga militar, estudiantil etc. - se pueden formar y añadir automáticamente al idioma. Estas palabras son básicamente nombres, verbos, adjetivos y adverbios las cuales pueden tener flexiones o desinencias y una variedad de formas derivadas que se conocen como "familia de palabras". La densidad léxica se puede calcular dividiendo el número total de palabras de contenido léxico entre el número total de palabras del texto oral o escrito.

Las palabras que no son palabras de contenido léxico se llaman *palabras funcionales o gramaticales*. Estas palabras incluyen pronombres, conjunciones, preposiciones, verbos auxiliares y algunos adverbios como por ejemplo "*then"*. Sirven para expresar relaciones gramaticales con otras palabras dentro de la misma frase o entre varias frases distintas o también para especificar la actitud o situación mental del hablante. Estas palabras decimos que pertenecen a una clase cerrada de palabras, ya que es muy difícil que se creen nuevas palabras funcionales. La importancia que tienen las palabras funcionales o las de contenido varía según los distintos idiomas.

c) *La sofisticación léxica*: también llamada "rareness" (Read, 2000: 203), es vocabulario altamente especializado, asociado con registros técnicos y argot y se supone que es bastante más difícil que el vocabulario menos especializado. La sofisticación léxica se calcula dividiendo el número de

familias de palabras sofisticadas en el texto por el número total de palabras del mismo.

d) *La Variación léxica* es la proporción de diferentes unidades léxicas en el texto. Una unidad léxica puede constar de más de una palabra, por ejemplo expresiones idiomáticas. La variación léxica se calcula dividiendo el número de diferentes unidades léxicas en el texto por el número total de unidades léxicas del mismo.

Otro índice que puede predecir la dificultad de un texto es la relación *type-token*, la cual relaciona el número de palabras diferentes usadas en un texto con el número total de palabras del texto.

La variación léxica, la sofisticación léxica y la densidad léxica nos informan de la riqueza léxica de un texto. Sin embargo, la característica esencial del nivel de competencia de los candidatos se demuestra en la habilidad de los mismos para reconocer y producir expresiones formadas por varias palabras. La longitud de las frases también nos informa de la complejidad sintáctica del texto oral o escrito.

1.13.5.2. Grado de contextualización

El grado de contexto en el texto nos indica también la mayor o menor dificultad a la que se enfrenta el candidato para interpretar su contenido. Es importante que un texto esté apoyado por una amplia gama de claves lingüísticas y paralingüísticas y que el lenguaje del texto se dé en un contexto familiar, conocido y que sea relevante para la información expresada en el discurso. Si un ejercicio de comprensión lectora contiene gran cantidad de información técnica o conceptos que no son familiares para los candidatos, es decir, si el conocimiento que presuponemos para el lector no existe o no es capaz de recordarlo, entonces el nivel de contextualización es muy reducido y el ejercicio será mucho más exigente a la hora de ser interpretado.

El grado de contextualización del discurso se halla al dividir la *información contextual* entre la *nueva información.* Un ejemplo de un test de comprensión lectora contextualizado sería un texto en el que el tema fuera familiar para el candidato.

1.13.5.3. La distribución de la información

La forma en la que esté distribuida la información es también importante. Un texto en el que la información sea muy compacta (nueva información distribuida en un espacio o tiempo relativamente pequeño) o muy difusa (nueva información distribuida sobre un espacio o tiempo muy largo) será más difícil de procesar y por lo tanto será más difícil llevar a cabo la tarea con éxito.

1.13.5.4. Tipo de información

El texto que contiene información concreta, positiva y que se atiene a los hechos objetivos será menos exigente y más fácil de procesar que una información abstracta, negativa o hipotética y que no se atiene a los hechos, por ejemplo oraciones condicionales. Cuanto más diferencia exista entre el mundo real y el hipotético planteado en el texto, más difícil será interpretar esa información. Así mismo cuanto más negativa sea la información contenida en el *input* más difícil será procesarla. Así entre la siguiente información:

a) I would like Mary to go to my party.

b) I wouldn't like Mary to go to my party.

c) I wouldn't like Mary not to go to my party.

La más fácil de procesar sería la primera, ya que la información es positiva y la más difícil la última por ser la información más negativa.

1.13.5.5. El formato

El formato en que se presenta la información incluye:

a) *La forma en que se presenta la misma* que puede ser de forma oral, escrita, en forma de dibujos, fotografías, tablas o gráficos que el candidato tiene que interpretar.

b) *El vehículo de presentación*. Si es un test de comprensión oral, por ejemplo, no es lo mismo oír a una persona presente en el lugar en el que se realiza el examen u oír la misma información grabada.

c) *El idioma en que se hacen las preguntas* o se dan las instrucciones. Esto es especialmente importante cuando se está examinando a un curso de nivel básico, ya que a veces las instrucciones son más difíciles de entender que el propio texto. De acuerdo con Sohamy (1984a), la investigación en los tests de comprensión lectora del inglés como lengua extranjera o EFL, sugiere que los candidatos obtienen mejores resultados cuando las preguntas sobre un texto de lectura en inglés se efectúan en su lengua materna.

d) *El grado de rapidez que se exige para llevar a cabo una tarea* como por ejemplo un ejercicio de comprensión lectora en el que se fija un tiempo máximo para localizar toda la información requerida, lo cual exige al candidato leer a gran velocidad, o un ejercicio escrito en el que se determina la extensión del mismo.

e) *El tipo de ejercicio que los candidatos tienen que hacer*. Hay tests cuyo formato, debido a las especificaciones de la tarea, es muy restrictivo, por ejemplo, un test de elección múltiple, ya que solamente hay que identificar o

seleccionar las respuestas. Otros tipos de tests son menos restrictivos como por ejemplo las pruebas de expresión escrita.

1.13.5.6. El tema

Si los temas sobre los que los candidatos tienen que hablar en un examen oral o leer en un examen de comprensión lectora o expresar sus opiniones en una tarea de expresión escrita son relevantes y familiares para el candidato, la ejecución de la tarea será mucho más fácil que si los temas no son relevantes. Así, se piensa frecuentemente que los exámenes de Proficiency de la Universidad de Cambridge no son adecuados para personas muy jóvenes, no porque no tengan conocimiento del idioma sino porque no tienen la experiencia y conocimiento del tema suficientes para expresar sus opiniones en un test de expresión escrita, por ejemplo. Cuando tenemos candidatos de diferentes culturas, edades y formación debemos evitar elegir temas que favorezcan solamente a parte de los mismos. Por ello es conveniente presentar varios textos distintos cuando se evalúa una destreza para intentar ser lo más imparcial posible,

1.13.5.7. El género y el registro del texto

El género y el registro del texto varían las características formales del mismo. La distribución de las frases, las formas pronominales y los párrafos, así como el orden de la información o los acontecimientos de una historia varían dependiendo de qué género y registro se considere. Por ejemplo, no es lo mismo leer o escribir un poema, una carta formal, una conferencia, un artículo para un periódico o un cuento.

Bachman considera que los distintos tipos de tests pueden ser considerados como ejemplos de distintos tipos de géneros y que el

desconocimiento de las características de un género determinado puede contribuir a que el test sea más difícil:

> It could be argued that particular types of language tests (for example, multiple-choice, cloze, dictation) themselves constitute genres, and that these activate certain expectations in test takers familiar with them, thus facilitating the task of test taking for these individuals, while making the task more difficult for test takers not familiar with the particular type. In addition, if the language of the input in a given test is characteristic of a genre that is unfamiliar to the test taker, we could hypothesize that tasks that depend on the interpretation of that input would be relatively difficult. Unfamiliarity with the characteristics of a given genre may also make the expected response more difficult. (Bachman, 1990: 138-139)

El saber utilizar el registro apropiado para distintas situaciones es esencial, por lo que en muchas ocasiones es necesario medir la habilidad del candidato para utilizar e interpretar más de un registro.

1.13.5.8. La variedad lingüística

No es justo que en el periodo de formación y práctica, los alumnos oigan y practiquen un dialecto o variedad de inglés determinado y a la hora de evaluarlos se les exija otra variedad. Convendría determinar primero cual es la variedad de inglés que los alumnos van a necesitar y después decidir en consecuencia, tanto a la hora del aprendizaje como de la evaluación. Si los alumnos han estado practicando con inglés Británico o Americano no se les puede evaluar con una muestra de inglés de la India por ejemplo.

1.13.5.9. La longitud del texto

Cuanto más largo es el texto, mayor es la influencia de todos los factores que se han indicado anteriormente y por lo tanto mayor es la dificultad. Para conseguir que un texto largo sea interpretable se necesita obligatoriamente que haya una buena organización del texto,

1.13.5.10. La organización del texto

La organización del texto depende principalmente de:

a) *La gramática* incluye un número de competencias, relativamente independientes, relacionadas con el uso de la lengua, tales como la morfología, la sintaxis, la fonología o la grafología, y el conocimiento del vocabulario que debe acomodarse al registro del texto. Estas competencias gobiernan la elección de palabras para expresar significados específicos, su forma, su distribución en frases para expresar ideas bien oralmente por medio de sonidos o bien de forma escrita por medio de símbolos.

b) *La cohesión* consiste en distintas formas de marcar explícitamente relaciones semánticas en el texto tales como elipsis, conjunción, cohesión léxica, referencia. También se ocupa de las convenciones que gobiernan la forma de ordenar tanto la información ya conocida como la nueva.

c) *La organización retórica* se ocupa de la estructura general del texto que debe seguir las convenciones ya establecidas según sea una narración, una descripción, un análisis, una carta formal, etc.

1.13.5.11. Las características pragmáticas

Las características pragmáticas de un texto contribuyen también a la dificultad del mismo. Van Dijk describe la importancia de la relación que existe entre las reglas del lenguaje y el contexto y los usuarios de un idioma. Las características pragmáticas de un texto forman parte de las condiciones que hacen que una expresión del lenguaje sea aceptable o no.

> Pragmatics must be assigned an empirical domain consisting of CONVENTIONAL RULES of language and manifestations of these in the production and interpretation of utterances. In particular, it should make an independent contribution to the analysis of the conditions that make utterances ACCEPTABLE *in some situation for speakers of the language*. (Van Dijk, 1977: 189-190)

Los candidatos tienen que conocer las convenciones sociolingüísticas que se utilizan para realizar determinadas funciones (petición, advertencia, consejo, amenaza, etc.) del idioma y las expresiones y palabras que son apropiadas y socialmente aceptables en un contexto determinado. Si estoy en una habitación cerrada, es invierno, tengo frío y la persona que está conmigo ha abierto la ventana y yo quiero que la cierre tengo dos opciones bien pedirlo directamente con una frase gramaticalmente correcta y con un claro significado de petición.

Ej. : " ¿Te importaría cerrar la ventana, por favor?"

o bien indirectamente utilizando una frase que generalmente no está asociada a la acción que yo quiero que se realice y cuyo significado depende mucho de las circunstancias y del contexto en el que se exprese la frase. Por ejemplo:

- "Me estoy quedando helada"
- "Hace un frío que pela"
- "Se estaba mejor antes"

En esta ocasión se pretende que la interpretación de la frase esté por encima de lo que se entendería utilizando únicamente la organización del texto de acuerdo a la gramática, la cohesión o la organización retórica.

1.14. Factores que afectan a los resultados de las pruebas de idiomas

1.14.1. Los correctores

La inconsistencia en la corrección puede ser un factor importante que afecte a los resultados de los tests de idiomas. Bachman y Palmer (1996), Bachman (2000), y Amengual (2004), han encontrado diferencias en el comportamiento de los correctores debido a diferentes factores tales como, la formación del corrector, la experiencia o la lengua materna. Sin embargo, ellos creen que los problemas de inconsistencia se pueden solventar preparando a los correctores, obteniendo un número suficiente de correcciones y estimando el grado de consistencia de las correcciones basándose en un número suficiente de pruebas (Bachman y Palmer, 1996: 227).

Además de todo lo anteriormente dicho, los que examinan deben ser conscientes de que cualquier cambio relacionado con la realización del test (organización, instrucciones, tiempo, etc.) puede conducir a cambios en la situación comunicativa y por lo tanto a cambios en los resultados del test. Ya hemos visto que el método de test o “facets”, como Bachman (1990) lo llama, influye en el comportamiento de un candidato y por lo tanto en su resultado final incluyendo las notas que provienen de los juicios que se realizan sobre la actuación del candidato. Sabemos que lingüística, psicolingüística y sociolingüística nos enseñan que debemos esperar actuaciones variables dependiendo de muchos factores diferentes.

Investigaciones sobre la evaluación de idiomas nos indican que los candidatos pueden llegar a una respuesta determinada a través de procesos muy diferentes. Si esto es así, resulta muy difícil determinar que conocimientos de una persona estamos evaluando con un test determinado y es igualmente difícil determinar, con cierto grado de certeza, lo que evalúa cualquier test. (Alderson, 1991a: 61).

Capítulo 2

PRINCIPALES CARACTERÍSTICAS DE LOS TESTS

2. 1. Introducción

Hemos visto que el tipo de examen que hagamos va a depender de las decisiones que se necesiten tomar. Puesto que las decisiones que se tomen van a afectar a las personas, debemos preocuparnos de la calidad de los resultados de nuestros tests. Cuanto más importante sea la decisión, en términos de impacto en individuos o programas, más debemos asegurarnos de que los resultados de nuestros tests sean fiables y válidos.

No existe un test o una técnica perfecta. Un test que puede ser ideal para un propósito puede no ser válido para otro. Una técnica que puede funcionar muy bien en una situación puede ser enteramente inapropiada para otra. Cada situación de evaluación es única y nuestra principal preocupación debe ser desarrollar tests que sean válidos y fiables, que tengan un efecto rebote o washback beneficioso y que sean prácticos. Bachman y Palmer (1996) afirman que no existen tests que sean buenos o malos en abstracto.

> … there is not such thing as a 'good' or 'bad' test in the abstract, and there is not such thing as the one 'best' test even for a specific situation. (Bachman y Palmer, 1996: 6)

Para ellos las características de un test son principalmente que tiene que ser útil y que pueda demostrarse de diferentes maneras que el idioma que utilizan los candidatos en el test se corresponde con el idioma que se utiliza en una situación distinta de la del examen.

Bachman y Palmer (1996: 18) explicitan *la Utilidad* con el siguiente esquema:

> **Usefulness** = Reliability + Construct validity +Authenticity + Interactiveness + Impact + Practicality

2.2. Fiabilidad

Un test completamente fiable sería uno en el que un individuo sacara siempre la misma nota si pudiera repetir el test varia veces. Así pues, la fiabilidad tiene que ver con la consistencia de las medidas tomadas a lo largo del tiempo, con pruebas de diferentes formatos corregidas por diferentes personas y otras circunstancias del entorno relacionadas con el test.

Fiabilidad es una cualidad esencial de los resultados de los tests, ya que a menos que las notas del test sean relativamente consistentes, no pueden proporcionarnos ninguna información sobre la habilidad que queremos medir. Bachman (1990: 160) dice que la fiabilidad es una condición necesaria para la validez, puesto que para que un test sea válido tiene que ser fiable. Sin embargo, él considera la validez como la característica más importante.

> We often think of reliability and validity as two distinct but related characteristics of tests scores. That is, although validity is the most important characteristic, reliability is a necessary condition to validity. When we consider the systematic effects of test method, however, we can see that this distinction may be somewhat blurred in language tests, since test method facets may affect both reliability and validity. (Bachman, 1990: 227)

Algunos investigadores (Brown, 1993a: 101) consideran la fiabilidad y la validez como dos características diferentes de los tests. Bachman (1990) los reconoce como aspectos complementarios de un factor común de medida. Hughes (1989: 42) considera que existe una gran interdependencia entre la fiabilidad y la validez, de forma que si hacemos que nuestro test sea más fiable, ello será a expensas de la validez. Al aumentar la fiabilidad disminuirá la validez y viceversa.

Otros investigadores, tales como Alderson et al. (1995), Alderson (1991a), Davies (1991) y Weir (2005a) creen que la fiabilidad es una forma de

validez y que a veces es muy difícil distinguir entre la una y la otra. Weir llama a la fiabilidad "*scoring validity*" refiriéndose a la medida en que los resultados de un candidato se mantienen estables en el tiempo:

> Scoring validity concerns the extent to which test results are *stable over time, consistent in terms of content sampling and free from bias*. In other words, it accounts for the degree to which examination marks are free from errors of measurement and therefore the extent to which they can be depended on for making decisions about the candidate. (Weir, 2005a: 23)

La fiabilidad de un test se puede cuantificar gracias al llamado *coeficiente de fiabilidad* o *índice de fiabilidad*. Las fórmulas más comunes son Kuder Richardson (KR 20), Kuder Richardson (KR 21) y el Alfa de Cronbach.

El coeficiente ideal de fiabilidad es 1. Un test con un coeficiente de fiabilidad de valor 1 daría exactamente los mismos resultados para un grupo dado de candidatos independientemente de cuando fuera administrado, y un test con un coeficiente de fiabilidad de valor cero daría resultados que no tendrían ninguna relación entre sí. Es decir, en este último caso los resultados serían diferentes dependiendo de cuando fuera administrado el test.

2.2.1. Formas de estimar la fiabilidad

La fiabilidad se puede medir de diferentes maneras:

2.2.1.1. La Fiabilidad test-retest.

Se usa para demostrar que un test está midiendo una habilidad determinada de forma consistente y se halla administrando el mismo test dos veces a un mismo grupo de sujetos. Después se correlaciona las calificaciones

obtenidas por las mismas personas obteniendo así un coeficiente de fiabilidad entre los pares de notas de los dos exámenes. Con este método se obtiene la consistencia de las calificaciones del test a lo largo del tiempo. El coeficiente puede variar entre -1 y +1 en una escala continua. Un valor de 1 indicaría una fiabilidad perfecta es decir, el test seria totalmente consistente. Por el contrario un coeficiente de valor cero indicaría una falta total de fiabilidad y por lo tanto el test seria totalmente inconsistente.

El problema con este método es que si el test lo hacen los candidatos dos veces consecutivas, la segunda vez que lo realizan pueden haberse acostumbrado al método o estar exhaustos. Por otra parte si permitimos intervalos más largos entre la administración de los dos exámenes, los candidatos pueden haber aprendido u olvidado conocimientos sobre el idioma. La memoria juega también una parte importante, habrá personas que recuerden parte del ejercicio y otras que no. Las correlaciones obtenidas pueden ser diferentes dependiendo pues no de la consistencia del test sino del tiempo transcurrido entre varias administraciones.

2.2.1.2. La fiabilidad de formas paralelas

La fiabilidad de Formas paralelas, también llamada *Parallel Form Reliability* o *Equivalent Form Reliability*, intenta solucionar los problemas que originaba el método de test-retest. Consiste en correlacionar las notas de dos tests equivalentes (parallel tests). Se crean dos versiones equivalentes o paralelas de una misma prueba para que sean realizadas por un mismo grupo y se halla la correlación de las calificaciones obtenidas por un mismo individuo. El coeficiente nos indicará hasta que punto una persona ha conseguido las mismas calificaciones con las dos versiones del test, y por lo tanto la influencia que tiene el contenido de los tests a la hora de evaluar a un mismo individuo.

Este método también es problemático, ya que es casi imposible elaborar dos tests genuinamente equivalentes pero no iguales.

2.2.1.3. La Fiabilidad de consistencia interna o internal consistency reliability.

Debido a los problemas asociados con los dos métodos anteriores, se introdujo el método de consistencia interna que analiza la consistencia u homogeneidad que existe entre las preguntas o los elementos internos de un test. Se puede estimar de varias formas pero la más común es el método de *split- half* o de análisis por mitades, que consiste en dividir una prueba única en dos mitades equivalentes. El candidato sólo realiza una prueba pero se le adjudican dos notas una por cada mitad del examen que después se van a comparar para ver la correlación entre ellas. Como se supone que todas las preguntas del test están midiendo la misma habilidad, los resultados de las dos mitades deberán ser los mismos. Cuanto más fuerte sea la correlación entre las dos mitades, mayor será la fiabilidad.

Weir nos advierte de que el método de análisis por mitades nos informa de la consistencia interna de un test en concreto pero no de la estabilidad temporal de los resultados:

> Split-half reliability offers a measure of consistency with regard to content sampling but obviously has nothing to say about the temporal stability of the scores as they result from a single administration of the test. (Weir, 2005a: 29)

De todos los métodos descritos anteriormente para hallar la fiabilidad de una prueba, el uso de coeficientes de consistencia interna es el más usado para estimar la fiabilidad de pruebas objetivas. El hecho de que estos coeficientes sean relativamente fáciles de calcular, por ejemplo, el Alfa de Cronbach o el KR20, ha originado que se consideren como el coeficiente más común o estándar, incluso por las agencias que se dedican a preparar distintos tipos de tests.

De todas formas los coeficientes de fiabilidad, como ya veremos más adelante, van a depender de muchos factores tales como la dificultad de la prueba, la longitud, la homogeneidad del grupo, etc. Ya veremos que si el grupo que hace el examen es muy homogéneo en sus conocimientos del idioma, no es muy probable que la consistencia interna de la prueba sea muy alta.

2.2.1.4. La fiabilidad entre correctores o inter-rater reliability

Es básicamente una variación de la fiabilidad de formas paralelas, ya que se calcula un coeficiente entre las notas que dan dos correctores. Este coeficiente sirve para cuantificar el nivel de acuerdo existente entre ambos correctores. Si el acuerdo es perfecto, la correlación será de 1, lo cual nos indicará que esos dos correctores darán siempre la misma nota cuando tengan que corregir una prueba de la misma calidad.

2.2.1.5. La fiabilidad de un mismo corrector o intra-rater reliability

Este tipo de fiabilidad está íntimamente relacionada con la fiabilidad del test-retest en el que se calcula un coeficiente entre las notas dadas por el mismo corrector a un mismo grupo de candidatos en dos ocasiones diferentes. Cuando evaluamos pruebas subjetivas, especialmente pruebas de expresión orales y de expresión escrita, la fiabilidad de los correctores es de suma importancia (Herrera, 2000 y Amengual, 2003, 2006).

No es sorprendente que una prueba tenga distinta calificación dependiendo de quien la haya corregido y todos conocemos en nuestros centros a personas que son consideradas como muy duras a la hora de calificar un examen y otras que son consideradas menos severas. Así pues, los comentarios que a veces se oyen a los alumnos diciendo que han tenido suerte

al aprobar un examen o que no han tenido suerte y han suspendido, desgraciadamente para nuestra profesión pueden ser ciertos. Incluso algunos profesores se sorprenden a veces cuando algunos alumnos suyos, a los que ellos conocen muy bien por haber estado en su clase todo el curso, suspenden o aprueban un examen que no ha sido corregido por ellos.

Estos casos aunque sean aislados tienen que dejar de existir y hay que conseguir que los correctores corrijan de forma similar y con los mismos criterios. Para ello es preciso realizar sesiones de estandarización para que todos los correctores lleguen a evaluar el mismo trabajo:

a) con la misma severidad
b) de forma consistente
c) teniendo en cuenta los mismos factores para todos los candidatos del mismo nivel.

Estos objetivos no se podrán conseguir si no se logra primero que cada uno de los correctores califique de forma consistente el mismo test u otro de calidad similar.

En un curso para profesores que realicé un verano en Gran Bretaña sobre Evaluación, se nos pidió que calificáramos individualmente un ejercicio escrito perteneciente a unos alumnos de un nivel determinado sobre el que habíamos estado trabajando esa semana. Así lo hicimos y, puesto que era viernes, nos fuimos de fin de semana después de que nos hubieran recogido el ejercicio que habíamos calificado.El lunes siguiente a última hora del día nos volvieron a dar los mismos ejercicios, otra vez sin corregir, para que los calificáramos de nuevo. Aún recuerdo el esfuerzo de los profesores para recordar qué calificación habíamos dado a cada ejercicio el viernes anterior y explicar el porqué los habíamos calificado así. Por supuesto pudimos comprobar después que la divergencia de notas entre los distintos profesores, y a veces entre las dos notas que un mismo profesor había dado a la misma prueba, era evidente e importante.

Yo solamente llevaba un curso como profesora y ese ejercicio me dejó marcada para el resto de mis días, ya que me hizo ver que por mucho que un examen sea válido, fiable y ajustado al nivel que se está evaluando, si la corrección no es fiable todo el trabajo anterior será inútil. Sin embargo, este es un aspecto al que no se le suele dar suficiente importancia en muchos centros y departamentos:

Weir resumía lo anteriormente dicho recalcando que un corrector, a la hora de calificar, debe de ser consistente consigo mismo y con los demás correctores.

> Markers need to be consistent in two ways: each marker has to be consistent within himself (intra-rater reliability), i.e., given a particular quality of performance, he needs to award the same mark whenever this quality appears, and there needs to be consistency of marking between markers (inter-rater reliability), i. e., one marker will award the same mark as another when confronted with a performance of the same quality.
> (Weir, 2005a: 34)

Hughes (2003) nos da una serie de consejos para aumentar la fiabilidad de la corrección a la hora de evaluar una prueba:

- Proporcionar una clave de corrección.
- Entrenar a los correctores.
- Identificar a los candidatos por número y no con su nombre.
- Que una prueba sea corregida por varios correctores independientes.
- Excluir las preguntas o ejercicios que no discriminan bien entre candidatos con distintos niveles de conocimiento del idioma.

- Ponerse de acuerdo en las respuestas que se van a aceptar como correctas antes de empezar la corrección.

2.2.2. Factores que afectan a los coeficientes de fiabilidad

El uso de coeficientes de fiabilidad para establecer la calidad de una prueba puede ser problemático, ya que los datos estadísticos pueden estar distorsionados debido a factores tales como:

2.2.2.1. La longitud de los tests

Varios investigadores, entre ellos Hughes, afirman que cuantas más preguntas tenga un test más fiable será:

> Other things being equal, the more items that you have on a test, the more reliable a test will be….. It has been demonstrated empirically that the addition of further (independent) items will make a test more reliable. (Hughes, 2003: 44)

Si asumimos que todas las preguntas o tareas de un test son indicadores representativos de la destreza que se quiere medir, cuantas más preguntas incluyamos más representativa será la muestra de esa destreza (Bachman, 1990).

2.2.2.2. La homogeneidad de conocimientos de los candidatos que hacen los tests.

Según Wood, hay que tener en cuenta si el grupo que realiza la prueba tiene un nivel de idioma muy homogéneo o no.

> ...since reliability being a correlation, is always influenced by range.... with a grater proportion of the ability distribution …, higher reliability estimates may be expected. (Wood, 1993: 138),

Si el rango de habilidad de los candidatos que hacen un test es muy amplio, el coeficiente de fiabilidad de ese test será mayor que si el grupo tiene un nivel muy homogéneo, ya que en este caso todas las notas serán muy parecidas, lo cual limitaría los coeficientes de fiabilidad de consistencia interna.

> Candidates of widely ranging ability are easier to rank reliably, and so will produce higher reliability indices than groups that are more equal in level where all the scores tend to bunch together lower standard deviation and lower variance. (Weir, 2005a: 32)

Por ello, al comparar coeficientes de fiabilidad de diferentes exámenes tenemos que tener en cuenta también a los candidatos que se presentan a esos exámenes. Si comparamos los coeficientes de fiabilidad de los exámenes de ESOL (English for Speakers of Other Languages) de Cambridge con los de TOEFL (Test of English as a Foreign Language), podríamos llegar a la conclusión de que los de Cambridge son menos fiables que los de ETS (Educational Testing Service) o los de IELTS (International English Language Testing System). Sin embargo, debido a la homogeneidad de los candidatos que se presentan a los exámenes de Cambridge, no es muy probable que sus coeficientes de consistencia interna puedan llegar a ser tan altos como los de TOEFL o los de IELST donde las personas que se presentan tienen amplia gama de conocimientos del idioma inglés.

2.2.2.3. La dificultad de los tests

Si el test es muy fácil o muy difícil para un grupo en particular, dará lugar a que el rango de calificaciones sea muy restringido y la varianza muy pequeña. Bachman (1990: 220). Esto es un problema a la hora de calcular la fiabilidad de los tests, especialmente si se trata de *norm-referenced tests,* ya que los valores de sus coeficientes son sensibles a las diferencias de resultados entre los distintos individuos. Por ello, cuanto mayor sea la discrepancia entre los resultados más fiables tienden a ser los tests.

Aunque, como ya veremos en el capítulo 3, no existe total acuerdo sobre si la comprensión auditiva y la comprensión lectora son destrezas unitarias o no (ver Alderson 2000a ó Weir 2005a), muchos consideran cierta la teoría que apoya que estas destrezas son parcialmente divisibles con lo cual los tests que miden los distintos aspectos de cada destreza no pueden ofrecer un alto grado de consistencia.

> If skills such as reading or listening are divisible, then high internal consistencies would not be expected in the papers testing these skills. (Weir, 2005a: 32)

2.2.2.4. La homogeneidad de las preguntas

Alderson y Clapham sugieren que la fiabilidad también depende de la homogeneidad de las preguntas. Si todos los ítems de un test miden el mismo aspecto de una destreza entonces las correlaciones entre ellos serán altas y viceversa.

> If all the items are intended to test the same skill in the same way, then the test items will intercorrelate highly, and the test will have a high reliability index. If the test contains sections testing different

> skills in different ways, these sections will not correlate highly with one another, and the reliability will be lower.
> (Alderson y Clapham, 1995: 88-89)

Un test que contenga un número alto de tareas tendrá un menor número de preguntas de la misma clase dentro de cada tarea, con lo que según lo explicado en el primer punto dará lugar a un coeficiente de fiabilidad bajo. De acuerdo con los dos últimos factores, la fiabilidad de un test depende no sólo de la consistencia interna de cada tarea sino que depende también de la variedad de tareas y del rango de habilidad de los candidatos.

Resumiendo, podemos aconsejar que cuando interpretemos un coeficiente de fiabilidad tendremos que tener en cuenta todo lo dicho anteriormente y recordar que la fiabilidad se ve siempre afectada por la naturaleza de los que hacen el test.

2.3. Validez

La consistencia es una cualidad deseable y necesaria para un buen test, pero también tiene que ser válida. La validez de un test se define como el grado en el que un test mide lo que dice estar midiendo. Es un procedimiento práctico para controlar el diseño, el desarrollo y el uso de los tests de idiomas. Weir (2005a) considera que incluso la fiabilidad, a la que llama "*score validity*", forma parte de la validez.

Henning define la validez como la cualidad que tienen los tests de medir lo que se supone que tiene que medir para un determinado propósito:

> Validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure. A test is said to be valid to the extent that it measures what it is supposed to measure. It follows that the term *valid* when

> used to describe a test should usually be accompanied by the preposition *for*. Any test then may be valid for some purposes, but not for others. (Henning, 1987: 89; en Alderson et al., 1995: 170)

La fiabilidad es una cualidad de las puntuaciones de los tests en si mismos, la cual nos indica hasta que punto los resultados no contienen ningún error de medida. La validez, por el contrario, es una cualidad de la interpretación y uso de los tests y nos informan si éstos son útiles, apropiados y significativos para un propósito en particular. Para demostrar la validez de un test, es necesario demostrar que correlaciona muy bien con índices de comportamiento que teóricamente cabía esperar que correlacionara, y que no correlaciona bien con variables que uno no espera que se correlacione.

2.3.1. Formas de medir la validez

La validez ha sido clasificada tradicionalmente en las siguientes clases:

2.3.1.1. Validez de contenido

La Validez de Contenido, como tradicionalmente se la describe, o "Context Validity" (Weir 2005a), nos indica si las tareas del test, que son una muestra de todas las posibles tareas con que puede evaluarse una destreza (estructuras, habilidades lingüísticas, etc.), son representativas o no. Esto incluye también las condiciones bajo las que se administra el test y el lugar en el que se administra.

Tenemos que asegurarnos que las preguntas o tareas del test cubren los aspectos más importantes de la destreza que intentamos evaluar y en la proporción correcta. Esto no está exento de problemas, ya que para poder asegurarnos de la validez de la muestra que incluimos en nuestros tests, debemos determinar con precisión el conocimiento que debe tener un alumno

de una determinada competencia del idioma. Así mismo, hay que conseguir, en la medida de lo posible, que el test obligue a reproducir comportamientos lo más parecidos posible a los de la vida real.

2.3.1.2. Validez criterial

La Validez Criterial *o* "Criterion-related Validity*"* se ocupa de hasta que punto existe una correlación entre las calificaciones del test y las calificaciones de otra prueba externa que se toma como referencia. Esta prueba externa tiene que ser adecuada, es decir, tiene que cumplir al menos las siguientes condiciones:

- Medir la misma habilidad que queremos medir con el nuevo test.
- Que se haya utilizado durante largo tiempo.
- Que sea ampliamente fiable y aceptada.

Una concordancia perfecta entre los resultados de los dos tests nos daría un coeficiente de valor 1 y la falta total de concordancia nos daría un coeficiente cero. Este tipo de correlaciones se suelen hacer cuando introducimos nuevas variables o especificaciones en un examen (variar el tiempo de examen, variar el número de preguntas, etc.) y queremos saber si el test resultante es todavía válido o no.

La validez criterial proporciona un criterio externo, el cual es también un indicador de la habilidad que se está evaluando (Bachman, 1990: 248).

2.3.1.2.1. Clases de validez criterial

La Validez Criterial se divide en *Validez Concurrente*, que indica que las dos pruebas fueron administradas en la misma época, y *Validez Predictiva*, que

nos informa de hasta qué punto el test puede predecir el comportamiento futuro de un candidato, por ejemplo, puede presagiar si un candidato que aprueba un examen de competencia va a ser capaz de trabajar en un determinado puesto, seguir un curso en la Universidad o un curso de idioma en una clase determinada. Esto es lo que hacemos con las pruebas de nivel con las que se intenta predecir la clase más adecuada para un alumno en particular.

El coeficiente de validez más usado dentro de la evaluación de idiomas es la validez concurrente que puede tener dos funciones:

1) Examinar las diferencias en el comportamiento de un test entre grupos de individuos de diferente nivel de idioma, por ejemplo, podemos analizar si un test discrimina entre candidatos nativos y no nativos. Asumiendo que los nativos son competentes en una serie de habilidades del idioma, debemos esperar que sus resultados sean notablemente mejores que los de los no nativos.

2) Examinar las correlaciones entre varias medidas de una habilidad del idioma.

De entre estas dos funciones la segunda es mucho más común que la primera. Los candidatos siempre están interesados en saber si un test está correlacionado con otro test más conocido. De ahí que muchos tests estándar o clásicos se administren junto con el test que queremos probar o introducir para saber si los resultados de los dos están correlacionados. Un problema que existe con este tipo de correlaciones es que dependiendo de la similitud de los tests la correlación puede ser interpretada como un factor de fiabilidad y no de validez.

La validez predictiva es problemática, ya que:

a) Compara resultados del test con otras medidas de los mismos candidatos recogidas tiempo después de realizado el test.

b) El éxito o el fracaso de una actividad depende de muchos factores aparte del conocimiento del idioma, por lo que puede haber variables que interfieran con la comparación realizada a lo largo del tiempo.

Hughes se pregunta hasta que punto tiene sentido usar los resultados finales como medida de criterio cuando hay tantos factores, aparte de la habilidad en la lengua, que contribuyen a ese resultado:

> How helpful is it to use final outcome as criterion measure when so many factors other than ability in English (such as subject knowledge, intelligence, motivation, health and happiness) will have contributed to every outcome? (Hughes, 1989: 25)

2.3.1.3. Validez de constructo.

Un test se dice que tiene validez de constructo si mide únicamente la habilidad que se supone que tiene que medir, por ejemplo la expresión oral o la comprensión escrita:

> In order to justify a particular score interpretation, we need to provide evidence that the test score reflects the area(s) of language ability we want to measure, and very little else. In order to provide such evidence, we must define the construct we want to measure. (Bachman y Palmer, 1996: 21)

Cuando se confecciona un test hay que asegurarse de que los constructos que estamos obteniendo, en la muestra de idioma que producen los alumnos al realizar ese test, son precisamente los que queremos evaluar y que no están contaminados por otras variables irrelevantes, tales como las características del método.

Si los constructos más importantes de un idioma no están representados en el test, el efecto rebote en la enseñanza que precede al test puede ser muy adverso, ya que los profesores pueden decidir no enseñarlos, puesto que no van a afectar a los resultados de las calificaciones de sus alumnos. Por ejemplo, si en un examen de idioma no evaluamos la destreza de expresión oral, existe una probabilidad muy alta de que esa destreza ni se enseñe ni se practique en clase, con lo cual causaríamos un perjuicio muy grande a nuestros alumnos. Ya hemos visto en este capítulo también que a veces las autoridades educativas cambian los constructos que se van a evaluar en un examen para forzar a los profesores a que los enseñen y a los alumnos a que los practiquen en clase.

La validez de constructo se utiliza para referirnos hasta qué punto se pueden considerar los resultados de un test como un indicador de la habilidad que queremos medir. La validación del constructo verifica empíricamente las hipótesis que especifican como deben comportarse los individuos en diferentes tests y que afirman que los candidatos deben comportarse de forma similar en todos los tests que midan la misma habilidad, mientras que el comportamiento en los tests que midan habilidades diferentes debe ser distinto.

2.3.1.4. Impacto

La validez que Bachman y Palmer (1996) llaman *Impacto*, Weir la describe como *Consequential validity.*

Ya hemos hablado extensamente en este capítulo de la influencia de los tests en la vida de muchas personas. Siempre que utilicemos un test tenemos que tener en cuenta que nuestra elección tendrá un impacto específico, tanto en los individuos como en la sociedad en general. Por lo tanto, los valores sociales y las consecuencias sociales no pueden ser ignorados cuando consideremos la validez de un test. “Tests have important effects on people’s lives and are thus potentially an instrument of power and control” (Weir, 2005a: 214).

El impacto de un test opera a dos niveles: En el ámbito de los individuos que están afectados por el test, principalmente profesores y alumnos, y a nivel del sistema educativo o de la sociedad.

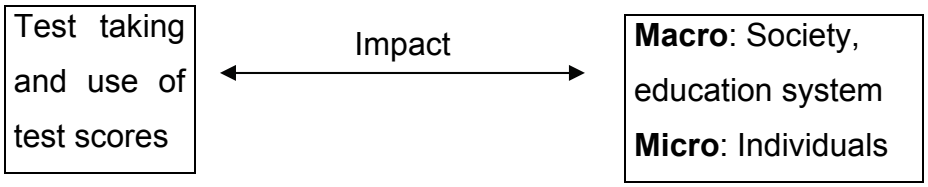

Figura 2.1. Representación del Impacto por Bachman y Palmer (1996: 30)

El analizar el impacto de un test sirve para identificar algunos de los efectos negativos de las prácticas actuales de evaluación sobre la motivación y la calidad de aprendizaje de los estudiantes e intentar neutralizarlos.

2.3.1.4.1. Efecto rebote, washback o backwash.

Bachman y Palmer (1996) creen que la noción de efecto rebote en la evaluación de idiomas se puede analizar en términos de impacto, ya que incluye el impacto potencial en los examinandos, en el sistema educativo, en las actividades de enseñanza y aprendizaje y en la sociedad:

> … includes the potential impact on test takers and their characteristics, on teaching and learning activities, and on educational systems and society. (Bachman y Palmer, 1996: 35)

Sin embargo, el concepto de efecto rebote para Hughes (1989: 45) es más restrictivo, ya que define backwash como ”the effect of testing on teaching and learning”, que puede ser beneficial o dañino, por ejemplo, cuando

aplicamos diferentes métodos de evaluación como: “direct testing” y ”criterion-referenced testing”.

Heaton (1990: 16) argumenta que los tests influyen en la enseñanza tanto positiva como negativamente, ya que los profesores, intentando ayudar a sus alumnos a aprobar los exámenes, ceñirán sus enseñanzas al tipo de examen que tengan. Sin embargo, Wall y Alderson (1995) nos dieron suficiente evidencia de que los tests tienen impacto sobre lo que los profesores enseñan pero no en como lo enseñan. Es decir, que los tests influyen sobre el contenido de un curso pero no sobre la metodología del profesor. También creen que el efecto rebote no es un tema simple y que “un test bueno” no garantiza un “efecto rebote bueno”.

Según Bachman (1990), el efecto rebote positivo ocurre cuando el examen utilizado refleja las habilidades y contenido enseñado en clase. Sin embargo, en muchos casos y particularmente en exámenes a gran escala, el currículo viene determinado por el examen, lo que conduce a un efecto rebote negativo. McNamara (2000) considera que las pruebas integradoras tienen un efecto rebote más positivo que las pruebas de elementos discretos. Shohamy (2000a) analiza la relación existente entre los exámenes de idiomas o “Language Testing” (LT) y el aprendizaje de un segundo idioma o “Second Language Acquisition” (SLA) y considera que esta relación es manifiestamente mejorable. Existe “limited relevance of LT to SLA”. Existe también la necesidad de ampliar el campo de evaluación de una lengua más allá de la psicometría involucrando a alumnos y candidatos, y teniendo en cuenta también los temas educativos:

> LT is to broaden its focus and scope by addressing broader views of language learning and language processing such as: viewing language and its complexities and dynamics; involving the learners and test takers; marketing better LT theories to those out of the field; expanding the context beyond psychometrics; expanding the types of instruments used beyond tests: addressing educational issues; and working towards relevance. (Shohamy, 2000a)

El efecto rebote también se describe en la literatura como manifiesto o encubierto. Según Prodomou (1995: 14) se considera que el efecto rebote es manifiesto cuando se utilizan en clase pruebas de exámenes o ejercicios del libro de texto que coinciden con los que los alumnos se van a encontrar en los exámenes. El efecto de este procedimiento es claramente negativo y se suele dar más importancia a la comprensión lectora y a la expresión escrita que a la expresión oral o a la comprensión oral. El efecto rebote encubierto es un proceso más inconsciente que proviene de la suposición de cómo aprenden los alumnos.

Este es un tema importante para Broadfoot (2005: 132), quien cree que, en el pasado, tanto los investigadores teóricos como los que elaboraban o desarrollaban los tests dieron más importancia al desarrollo y a la administración de exámenes que al impacto de los mismos en los estudiantes que se examinaban. También afirma que ha habido un olvido colectivo del efecto rebote y del impacto que tienen los tests.

2.3.1.5. Validez aparente

La Validez Aparente o "Face Validity" refleja la medida en la que el test es atractivo para los que se examinan y por lo tanto afecta a la credibilidad y la aceptabilidad del test por parte de las personas que no son expertas en la materia, por ejemplo, alumnos, padres o administradores del test. Este término es desechado por Low (1985: 157) debido a que tiene una connotación peyorativa. Él sugiere agrupar todas las clases de validación que implican juicios subjetivos, bien sean de expertos o de profanos en la materia o bien de los que usan o de los que elaboran los tests, y poner a la Validez Aparente otra etiqueta con tonos menos despectivos, por ejemplo, *Validez Percibida*. Stevenson (1985) afirma que la validez aparente puede ser engañosa: "face validity judgements are naïve because appearances in testing are treacherous,

and well-established deceivers", y añade que los mejores tests son los que están basados en estudios de validez y fiabilidad y no en las apariencias:

> By common professional agreement – through set standards and ethics - we believe that the best tests are those which specify theoretical bases an assumptions, which report reliability and validity studies and data, which emphasises the "ifs" and the "buts", which take care to issue warnings and doubts. In short, the best tests are those which do not push a product through packaging, and do not promise to do what no test can do. As a result, however, popular face-validity judgements will always tend to favour the worst tests - those that tell the gullible that they do what no test can do, and have the looks to "prove" it.
> (Stevenson, 1985: 114)

Bachman y Palmer (1996: 42) y Bachman (1990: 285) no consideran la apariencia de validez una cualidad separada del resto de cualidades del test. De forma semejante Alderson et al. (1995: 173) creen que la validez aparente es importante porque si los candidatos aceptan y consideran el test como una prueba válida, puede que se lo tomen más seriamente y es más probable que actúen lo mejor que sepan dentro de cada habilidad. Es decir, cree que la validez aparente afectará a la validez de la repuesta.

El fondo del asunto, desde el punto de vista práctico, es si los candidatos y las instituciones que utilizan los tests los aceptan y los consideran útiles. Por esta razón la apariencia de un test es muy importante, aunque una vez reconocido esto a veces no sepamos que tipo de tarea va a ser aceptable y cual no. Por todo lo expuesto anteriormente, y aceptando que la apariencia de los tests afecta a la aceptabilidad de los mismos por parte de los candidatos y de las instituciones que los utilizan, es importante que un test de lenguaje comunicativo tenga la apariencia de estar realizando con el idioma una tarea semejante a algo que haríamos "en el mundo real".

Hughes resume todo lo expresado sobre este tipo de validez cuando afirma que aunque la validez aparente no es un concepto científico, sin

embargo, esta cualidad es muy importante, ya que sin ella puede que el test no sea usado por las autoridades, instituciones o empresarios:

> Face validity is hardly a scientific concept, yet it is very important. A test which does not have face validity may not be accepted by candidates, teachers, education authorities or employers. It may simply not be used; and if it is used, the candidates´ reaction to it may mean that they no not perform on it in a way that truly reflects their ability. (Hughes, 1989: 27)

2.3.1.6. Validez de la respuesta

La validez de la respuesta está relacionada con la recogida de la información. Se preocupa de cómo los candidatos responden a las preguntas y explora hasta que punto las respuestas revelan los procesos y los resultados que las personas que elaboraron el test querían que se produjeran. También analiza si una misma prueba puede dar lugar a que diferentes candidatos realicen diferentes procesos mentales para resolverla y por lo tanto originaría que una misma prueba estuviera midiendo aspectos diferentes según el candidato.

> The process they go through, the reasoning they engage in when responding, are important indications of what the test is testing, at least for those individuals. (Alderson et al., 1995: 177)

2.3.2. Importancia de la validez en la investigación

La importancia de la validez en la investigación de la evaluación del lenguaje ha sido subrayada por un amplio número de investigadores en los últimos años. Ej. Alderson (1991a), Davies (1991), Cumming and Berwick

(1996), Weir (2005a) o Bachman (2000: 23) quien afirma que: “validation has become the facto paradigm for language testing research and development”.

Davies (2003) considera que la ortodoxia en la evaluación de idiomas está en el mantenimiento del equilibrio entre la fiabilidad y la validez, y que en los últimos años se han cometido ciertos abusos en un intento por promover y explorar la validez. Esto ha originado ciertos desequilibrios relacionados con la definición del constructo del lenguaje. Estos desequilibrios, a los que él llama herejías, los resume en tres puntos:

1. - La herejía de la lengua: el punto de vista de que la lengua no es sólo estructura sino también uso y variedad ha llevado a que nos olvidemos de la estructura para considerar solamente el aspecto del uso de la lengua.

2. - La herejía de la evaluación: la asunción de que debido a que los tests tienen un impacto, su efecto rebote incluye todos los efectos correlacionados, no solamente los que pueden ser considerados como casuales.

3. - La herejía de la investigación y el desarrollo: la incorporación a la investigación de la evaluación de idiomas de nuevos sistemas de expresión y nuevos métodos de análisis que amenazan con convertirse en dominantes, determinando que evaluar y como juzgar los tests. Por ejemplo, el uso de ordenadores para evaluar idiomas.

Podemos resumir diciendo que la validez no es una característica de un test sino que está relacionada con la interpretación de los tests y sus resultados, asumiendo que los usos e interpretaciones tienen consecuencias. La validez es un proceso mas que una posesión, pero los que elaboran un test tienen que comprometerse cuando lo recomiendan a dar toda la información que posean sobre la validez del mismo, sabiendo y haciendo la observación de que pueden estar equivocados.

2.4. Autenticidad

También tenemos que asegurarnos que el contexto sea lo más relevante y parecido posible a una situación en la vida real. La preocupación por la autenticidad en el diseño de un test ha conducido a que se utilicen una gran variedad de tareas en la evaluación de los idiomas. Tests descritos como directos, funcionales, comunicativos, unificados, de tareas de la vida real, de producción, etc., se conocen últimamente como tests auténticos.

Bachman (1990: 307) define la autenticidad como: "the extent to which test tasks replicate 'real life' language use tasks". Tiene que existir relación entre el lenguaje usado en los tests y el lenguaje usado en la vida real, ya que si esa relación no existe, puede que nuestros tests de idiomas no nos digan nada sobre la habilidad que deseamos medir (p.356). Sin embargo, reconoce la complejidad del tema, ya que una tarea puede tener un grado alto de autenticidad en la situación y muy bajo en la interacción o viceversa. En otras palabras, el test no es necesariamente auténtico o no auténtico sino que depende de si la tarea de evaluación guarda alguna relación con el contexto real en el que se llevaría a cabo:

> Tasks would not necessarily be either authentic or non-authentic, but would lie on a continuum which would be determined by the extent to which the assessment task related to the context in which it would be normally performed in real life.
> (Bachman, 1990)

Bachman y Palmer (1996:23) consideran de gran importancia la autenticidad de un test, ya que es lo que nos permite generalizar a la hora de interpretar las calificaciones. Para ellos la autenticidad es el grado de correspondencia de las características de un test de idioma con las características de la lengua que se utilizaría para llevar a cabo la tarea en la

vida real: “the degree of correspondence of the characteristics of a given language test to the features of a TLU (target language use) task”.

Esta relación se muestra en la siguiente figura:

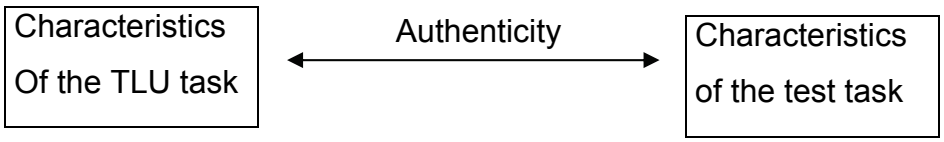

Figura 2.2. Representación de la Autenticidad por Bachman y Palmer (1996: 23)

Cuanto más grande sea la semejanza, con mayor confianza podremos relacionar la respuesta de un individuo al test con su probable comportamiento en una situación de la vida real. También afirman que la autenticidad tiene un potencial efecto sobre la actuación del individuo.

Por otra parte, Lewkowicz (2000), que estudió la importancia de la autenticidad de los tests para los candidatos, llegó a la conclusión, basándose en los resultados de su estudio, de que la autenticidad puede ser de capital importancia para los investigadores y los que elaboran tests, ya que teóricamente es la forma que tienen para poder generalizar y aplicar los resultados de una situación de test a otra situación de la vida real. Sin embargo, esta característica no es tan importante para los candidatos u otras personas involucradas en el proceso de evaluación.

Spence-Brown (2001) llega a la conclusión de que el mismo acto de examinar cambia la naturaleza de una tarea potencialmente auténtica. Ella concluye que la autenticidad debe estar relacionada con la realización de una actividad, no con su diseño.

Algunos exámenes de los más conocidos, como pueden ser los exámenes de Cambridge, han cambiado algunas de sus pruebas, por ejemplo,

las de elección múltiple, para proporcionar mayor autenticidad a los exámenes. Lo mismo se ha venido haciendo en las pruebas de Certificado de las Escuelas Oficiales de Idiomas de la Comunidad de Madrid. Aunque esto puede acarrear alguna consecuencia negativa, ya que al disminuir el número de preguntas, debido a la limitación del tiempo de los exámenes, la consistencia interna de la prueba puede disminuir también. Sin embargo, no tenemos que olvidar que la validez de un test depende además de otras muchas características del mismo.

2.5. Carácter interactivo

Bachman y Palmer definen *interactiveness* como la forma en que el candidato se involucra en la tarea para cumplirla:

> …the extent and type of involvement of the test taker´s individual characteristics in accomplishing a test task… . The interactiveness of a given language test task can thus be characterized in terms of the ways in which the tests taker's areas of language knowledge, metacognitive strategies, topical knowledge, and affective schemata are engaged by the test task.
> (Bachman y Palmer, 1996: 25-29)

Como en un examen de idioma lo que queremos evaluar es el conocimiento que un candidato tiene del mismo, tenemos que diseñar tareas en las cuales el individuo tenga que hacer uso de ese conocimiento del idioma para interactuar con el material proporcionado por el test. Si la interacción de un candidato con el test no requiere el uso de la lengua, entonces no podremos, basándonos en su actuación, hacer inferencias acerca de su destreza de la lengua.

Tanto la autenticidad como el carácter interactivo de los tests son relativos. Podemos crear un test que sea auténtico y de carácter interactivo para un grupo de candidatos y no para otro, ya que los candidatos pueden

procesar la misma tarea de forma diferente. Por otra parte, los niveles mínimos aceptables de autenticidad y carácter interactivo deben estar en equilibro con el nivel del resto de cualidades del test.

2.6. Factibilidad

La factibilidad se puede definir como la relación entre los recursos que se requerirán en el diseño, desarrollo y uso de los tests y los recursos disponibles para esas actividades. Bachman y Palmer (1996) representan la factibilidad, que ellos llaman “Practicality”, con la siguiente fórmula.

$$\text{Factibilidad} = \frac{\text{Recursos disponibles}}{\text{Recursos necesarios}}$$

Si la factibilidad es ≥ 1, el desarrollo y uso del test es práctico.
Si la factibilidad es ≤ 1, el desarrollo y uso del test no es práctico.

Figura 2.3. Representación de la Factibilidad por Bachman y Palmer (1996: 36)

Por lo tanto, un test será factible o práctico si su diseño, desarrollo y uso no requieren más recursos que los que disponemos. Weir no incluye la factibilidad dentro de las características necesarias para que un test sea válido. Él afirma que simplemente no la considera una condición necesaria para la validez.

> Only when sufficient validity evidence is available to justify interpreting test scores as an acceptable indication of the control of

> an underlying construct should we concern ourselves with practicality. If practicality is allowed to intrude before such evidence is available we run the risk of not assessing what we want to.
> (Weir, 2005a: 49)

2.7. Conclusiones

Es necesario pues, tener muy claro lo que deseamos evaluar para después elegir el método más adecuado sin que interfiera ninguna consideración práctica que pueda amenazar la validez y fiabilidad de un test. Si después de analizar todas las características expuestas anteriormente como factores que pueden afectar a los resultados de los tests, se comprueba que grupos diferentes de individuos, que supuestamente tienen el mismo nivel de idioma en una destreza, sistemáticamente obtienen resultados diferentes, tenemos que considerar la posibilidad de que los tests estén sesgados.

Para hallar las posibles causas del sesgo tendremos que considerar la formación de los candidatos, sus antecedentes culturales, su conocimiento previo del tema, su idioma materno, la ambigüedad de las preguntas, si las tareas miden habilidades independientes, el género o la edad como características que pueden originar que el test esté sesgado a favor o en contra de algunos candidatos.

Además la evaluación de un idioma ocurre en una sociedad determinada perteneciente a una clase social dada que posee una educación y un sistema educativo definidos. Como ya se ha expuesto en este capítulo, el uso de los tests de idiomas está determinado mayormente por las necesidades políticas, las cuales dependen del tipo de sociedad y de la época. Por todo esto, aparte de las teorías psicométricas y la lingüística aplicada tenemos que considerar para qué finalidad política o social se usan los tests que elaboramos y si son útiles para la función para la que fueron creados. Sin olvidar, por supuesto, el efecto rebote y las consecuencias positivas o negativas que nuestros tests

pueden tener sobre el sistema educativo o sobre los individuos que realizan los tests.

Shohamy (1984) critica que la mayoría de los enfoques que se utilizan actualmente en la elaboración de exámenes de idiomas se centran en uno de los siguientes aspectos:

a) La definición de idioma y lo que significa conocer un idioma.

b) Todos los criterios que se utilizan para indicar que el test es un buen instrumento para juzgar y tomar decisiones sobre el conocimiento de la lengua de un individuo. Estos criterios son principalmente la fiabilidad y la validez.

Igualmente, Shohamy aboga por la integración de estos dos componentes para producir buenos tests:

> Tests which are based on language learning theories without good psychometric theories are not sufficient in the same way that tests which possess measurement properties but are not based on complete language learning theories are insufficient.
> (Shohamy, 1984: 161)

Solamente teniendo en cuenta el aprendizaje de la lengua y las teorías psicométricas, podremos elaborar tests que discriminen entre una amplia variedad de candidatos, que sus elementos se ajusten al constructo que se quiere evaluar, que produzcan resultados similares en una segunda administración, y que sean lo suficientemente difíciles para los buenos alumnos y suficientemente fáciles para los alumnos de niveles inferiores.

Por otra parte, cuando elaboremos un test, hay que tener en cuenta que actualmente además de los aspectos lingüísticos, es necesario considerar también los aspectos sociolingüísticos del idioma. Hoy todo el mundo ha asumido que conocer un idioma significa también saber “como”, “cuando”,

“donde”, y “con quien” utilizarlo. Es decir que toda esta información tiene que formar parte también de un test de competencia sociolingüística.

Todos los factores que hemos señalado como importantes para diseñar un test que sea válido han sido tenidos en cuenta a la hora de confeccionar el C-test que hemos utilizado en nuestro estudio para analizar si podía formar parte de la batería de pruebas que se utilizan para evaluar la destreza de comprensión lectora en las Escuelas Oficiales de Idiomas de la Comunidad de Madrid.

Capítulo 3

EVALUANDO LA COMPRENSIÓN LECTORA

3.1. Introducción

La evaluación de la habilidad de la comprensión lectora es de gran importancia en una amplia gama de escenarios educativos y profesionales. Es la destreza de idiomas sobre la que más se ha investigado y a la vez la más enigmática, ya que su proceso es generalmente silencioso, interno, y privado. La mayoría de los investigadores están de acuerdo en que la lectura tiene las características de ser una actividad rápida, interactiva y con un propósito. Alderson (2000a). El leer es interactivo puesto que el texto no posee un "significado" que está esperando ser descubierto por un hábil lector sino que los conocimientos, cualidades y experiencias que posee el lector interaccionan con el contenido del texto creando significado. Por lo tanto, tanto las características del lector como las del texto afectan al proceso de lectura.

También están de acuerdo en que existen distintos niveles de comprensión, ya que es posible entender las palabras pero no entender el mensaje de una frase, o entender las frases y no entender la organización del testo. De ahí que los procesos que el candidato realiza para comprender un texto se dividan en microprocesos, que son los relacionados con la comprensión local o frase a frase, y macroprocesos, que son los relacionados con la comprensión global del texto.

Pulido define la lectura como una actividad cognitiva compleja que implica diversos y simultáneos procesos lingüísticos:

> A complex cognitive activity, involving simultaneous linguistic processing such as pattern recognition, lexical access, concept activation, syntactic analysis, propositional encoding, sentence comprehension, and intersentence integration, as well as the activation of prior knowledge, information storage, and comprehension monitoring. (Pulido, 2007: 155)

La competencia lectora de una persona dependerá de la eficiencia con que lleva a cabo los diversos procesos anteriormente mencionados. Por todo esto, los evaluadores cuando se enfrentan a la tarea de construir un test de comprensión lectora saben de antemano que su forma de entender el constructo es defectuosa, parcial y difícilmente perfeccionable.

El leer con un propósito proporciona motivación, lo cual es un aspecto muy importante del hecho de ser un buen lector. El propósito de la lectura determinará las operaciones que se realicen en el acto de leer o las estrategias que se utilicen. El lector puede tener que obtener bien una información literal del texto, bien tener que inferir esa información, o quizás realizar una evaluación crítica del texto. Esta es la razón por la que los motivos para leer, proporcionados en los tests de nuestros exámenes, deben ser lo más realistas posibles, los textos deben ser interesantes y las tareas deben asemejarse a las que los candidatos se van a encontrar en sus vidas

> Reading might be tested within a content-focused battery: texts that carry meaning for readers, that interest them, that relate to their academic background, leisure interests, intellectual level and so on, might motivate a deeper reading than the traditional, relatively anodyne or even contentless texts.(Alderson, 2000a: 29)

Tenemos que tener también en cuenta la dificultad o *legibilidad* del texto y sabemos que los textos con gramática más compleja, con contenido menos familiar, con vocabulario de menor frecuencia, con alta densidad léxica o con expresiones idiomáticas, son más difíciles de entender por el lector. Sin embargo, los lectores deben conocer como se organizan los textos. Por ejemplo, donde se encuentra la idea principal en un párrafo o como se indican los cambios de contenido.

De entre todos los factores anteriormente mencionados, han sido la gramática y el vocabulario sobre los que ha recaído la mayor atención de los investigadores. Los estudios que se han realizado demuestran que ambos pueden ser utilizados como predictores de la habilidad de la comprensión

lectora. (Unquhart & Weir, 1998; Shiotsu & Weir, 2007). Barnett (1986) estudió la influencia de la gramática y el vocabulario por separado y llegó a la conclusión de que tanto el conocimiento sintáctico como el de vocabulario afectan a la comprensión lectora, ya que un aumento de los mismos en sus estudiantes se refleja de forma casi simétrica en los resultados de la comprensión lectora.

Aunque existe la idea generalizada de que el vocabulario es uno de los factores que más influyen en la dificultad de un texto, y que más relación tiene con la habilidad de la lectura, sin embargo, la investigación ha demostrado que es la gramática la que más contribuye a la comprensión de un texto. Shiotsu & Weir (2007: 118) demuestran que el conocimiento sintáctico de un candidato predice mejor su actuación en un test de comprensión lectora que su conocimiento de vocabulario. Alderson (2000a) también afirma que existen correlaciones muy altas entre los tests de gramática y los diferentes tests de lectura y que a veces la correlación es más alta entre la gramática y los distintos tests de comprensión lectora que entre estos tests de comprensión lectora entre sí. Con esta investigación se demuestra que no es necesario hacer un examen de gramática separado del de lectura en la batería de tests de competencia.

3.2. Destrezas y estrategias en que se divide la habilidad de la comprensión lectora

Todavía no hay acuerdo sobre si la lectura es divisible entre las distintas habilidades o destrezas de las que se compone, las cuales pueden ser identificadas claramente, o si se trata de un proceso unitario. Algunos investigadores afirman que los estudios empíricos han proporcionado evidencia contradictoria sobre si las distintas destrezas en que se ha dividido la lectura son realmente diferentes (Alderson, 2000a).

En esta situación, si queremos evaluar la comprensión lectora como un constructo diferente de la expresión escrita o la comprensión oral, nos vemos forzados a descomponer la lectura entre lo que creemos que son sus componentes.

- *Skimming o lectura superficial* – leer por encima para captar la idea principal del texto. Este tipo de lectura es selectivo, algunas secciones del texto se pueden omitir o prestarles muy poca atención. Se construye una idea general con el menor número de detalles del texto posibles.

- *Scanning o lectura rápida* – leer rápida y selectivamente para conseguir objetivos muy concretos. Por ejemplo, encontrar un número en una guía de teléfonos, localizar nombres, fechas, etc.

- *Lectura cuidadosa* – está asociada con la actividad de leer para aprender y por lo tanto con la lectura de libros de texto. El proceso no es selectivo, ya que el lector intenta manejar y absorber la mayor parte de la información contenida en el texto.

- *Búsqueda* – se trata de localizar información sobre temas predeterminados. Por ejemplo, información para contestar a unas preguntas, completar una tabla, etc. La búsqueda de información está guiada de antemano y no es necesario captar la idea principal del texto.
  (Urquhart & Weir 1998; en Weir, 2005a: 90).

Por su parte Grabe (1991) propone que cualquier proceso de lectura fluida se compone de los siguientes seis elementos:

- Automatic recognition skills
- Vocabulary and structural knowledge

- Formal discourse structural knowledge
- Content / word background knowledge
- Synthesis and evaluation skills / strategies
- Metacognitive knowledge and skills monitoring

Entre los elementos cognitivos incluye: reconocer la información más importante del texto, ajustar la velocidad de lectura, prever, usar el contexto para resolver un malentendido, formular preguntas sobre la información, leer el texto superficialmente, etc.

Si podemos identificar destrezas y estrategias que parecen contribuir de forma importante al proceso de la lectura, debería ser posible evaluarlas y usar los resultados globales para decidir sobre la competencia en la comprensión lectora de un candidato. Sin embargo, todavía no se sabe muy bien si estas destrezas o estrategias que hemos descrito por separado interaccionan entre sí, o que relación existe entre ellas, o si alguna de estas destrezas incluye a otras, por lo que se admite la posibilidad de que la suma de las partes puede no equivaler completamente a lo que los lectores han comprendido del texto.

Cuando realizamos un examen a nuestros alumnos, lo que importa no es el saber si aprueban o no el test sino hasta que punto los resultados obtenidos en el test pueden generalizarse para determinar su habilidad de lectura en el mundo real. Si nuestros tests están basados en una teoría de lo que implica la habilidad de leer, entonces la generalización de los resultados de nuestros tests será más fácil. Como complemento al enfoque basado en una teoría, Alderson (2000a) propone que las tareas de los tests que diseñamos tienen que ser semejantes a las actividades de lectura que realizamos en el mundo real. Después tenemos que considerar que destrezas se requieren para completarlas con éxito y así definir el constructo que esa tarea está midiendo.

3.3. Factores que pueden afectar a la comprensión lectora

Entre las variables que afectan a dificultad del texto se encuentran el vocabulario, la complejidad sintáctica, la legibilidad, la cohesión, la coherencia y el tema. Ya hemos hablado de algunas de ellas en la introducción y otras van a ser analizadas con más detenimiento.

3.3.1. El conocimiento del mundo

También se ha mencionado que para poder procesar una lengua se requiere cierto conocimiento del mundo. La activación de este conocimiento es rápida y automática, y sin esos procesos, la comprensión, si es que tiene lugar, es lenta y laboriosa. Por ello, el conocimiento del mundo es necesario para la comprensión lectora.

3.3.2. El propósito y la motivación del lector

Según el propósito de la lectura el lector va a estar más o menos motivado y esto va a influir sobre la calidad de la lectura. Fransson (1984) demostró que los alumnos que estaban extrínsecamente motivados porque esperaban un test después de la lectura, por ejemplo, leían superficialmente, prestando más atención a los detalles que a las ideas principales, o a como se relacionaban las ideas del texto entre sí, o a lo que el alumno conocía sobre el tema o sobre el mundo.

Así pues, cuando los alumnos encuentran una situación de amenaza la comprensión del texto es de menor nivel que la que podrían obtener en otra situación diferente. Esto podría indicar que los datos obtenidos con lecturas informales podrían ser cualitativamente mejores que los obtenidos en los tests

de lectura. A todo lo dicho anteriormente, podemos añadir el *estado emocional* del candidato en un examen, ya que es sabido que el estado de ansiedad de los examinandos también afecta negativamente a sus resultados (Fransson, 1984).

3.3.3. El contenido

Como ya se ha dicho, es una asunción general, tanto en la primera lengua como en la segunda, que la lectura es un proceso de interacción entre el lector y el texto, y que los temas conocidos pueden hacer de puente entre ambos, ya que proporcionan conocimiento anterior o "*content squemata*" (Carrell, 1987b). De esta forma, los temas familiares pueden facilitar la comprensión del lector que prestaría menos atención al significado y más atención a la forma y estructuras del texto (Pritchard, 1990).

Lee (2007) llevó a cabo un proyecto de investigación para comprobar la veracidad de estas teorías, y sus resultados revelaron que la familiaridad del tema ayudaba a la comprensión del texto pero no era efectiva para facilitar el aprendizaje de estructuras. Por otra parte, el hacer que los alumnos prestaran más atención a la forma y estructuras del texto perjudicaba a la comprensión del significado de la lectura. La tarea será más o menos exigente dependiendo del tema del texto, del tipo de tarea, y sobre todo de los conocimientos culturales y del mundo que posea el lector:

> Reading is an interaction between a reader with all that the reader brings with him/her – background knowledge, affect, reading purpose, intelligence, first language abilities and more – and the text, whose characteristics include topic, genre, structure, language (organization, syntax vocabulary, cohesion) and so on.
> (Alderson y Banerjee, 2002: 84)

Sin embargo, varios estudios muestran que una competencia lingüística superior del idioma puede compensar la falta de conocimiento del tema, y que la familiaridad con el tema puede compensar la menor competencia lingüística (Alderson, 2000a: 44).

De hecho, nos encontramos con estudios cuyos resultados parecen contradictorios: Perkins & Brutten (1988b) y Hale (1988) llegaron a la conclusión de que los estudiantes universitarios obtienen mejores resultados cuando los textos están relacionados con sus disciplinas, con lo cual los evaluadores deben controlar los temas de los tests de comprensión lectora. Sin embargo, Alderson & Urquarth (1984, 1985) y Clapham (2000) demostraron que los alumnos no obtienen necesariamente mejores resultados cuando los exámenes contienen textos relativos a sus materias, y de hecho Clapham sugiere sustituir los textos específicos de los tests de comprensión lectora del examen de EAP (English for Academic Purposes) por tests de conocimiento gramatical y aptitud académica.

Sasaki (2000) examinó los efectos de los esquemas culturales en los procesos de examen con la técnica de *cloze* y utilizó inmediata retrospección para obtener los datos. A los alumnos de EFL se les dio dos versiones: una que era culturalmente familiar y otra que no, y se analizó como contestaron a ambas versiones. Los resultados mostraron que los alumnos que leyeron el *cloze* cuyo texto les era culturalmente familiar intentaron resolver más ítems, y generalmente entendieron el texto mejor que los candidatos que leyeron el texto original cuya cultura no reflejaba la suya propia. Esto demuestra que no solamente el tema puede afectar a los resultados de un test sino que también los esquemas culturales del texto pueden ser importantes. También demuestra, que los alumnos que tuvieron mejores resultados por haber leído el texto que les era familiar hicieron más uso de la información dentro de la frase o “intra-frase” que fuera de ella o ”interfrase”.

Un problema al que se enfrentan los estudios sobre el conocimiento del contenido y conocimiento cultural es distinguir entre conocimiento cultural y del contenido y conocimiento de vocabulario. Esto es crucial para los que elaboran

tests de comprensión lectora como segunda lengua que quieren evitar sesgos debido al contenido del texto, pero que aceptan que la falta de vocabulario relevante para los candidatos origina peores resultados en los tests.

3.3.4. El sexo

Otra idea muy generalizada es que los resultados de la evaluación de la comprensión lectora pueden variar considerablemente de acuerdo con el contenido de los textos y el género de los candidatos, creyendo que los pasajes de comprensión de lectura de contenido científico favorecían a los candidatos del sexo masculino y los pasajes con temas no científicos favorecían a los del sexo femenino. Sin embargo, una investigación realizada por Pae (2003) con estudiantes coreanos reveló que no existe evidencia de que el contenido de los textos de la comprensión lectora favorezca a uno o a otro sexo, o que al menos que esto no ocurre con todas las nacionalidades.

3.3.5. Nivel de competencia de la lengua

Otro problema que tenemos que resolver y que todavía no está claro es, cuando evaluamos comprensión lectora en un segundo idioma, si estamos midiendo la competencia y las habilidades lingüísticas del idioma o la habilidad de la comprensión lectora. Necesitamos conocer cual es la diferencia entre leer en un segundo idioma y conocer el idioma. No estamos seguros de si los alumnos pueden transferir alguna de las destrezas de comprensión global que poseen en su primer idioma, lo cual tendría implicaciones importantes a la hora de evaluar la comprensión lectora.

El tema de si leer en un segundo idioma es un problema de lengua o un problema de lectura es todavía objeto de investigación. (Alderson, 1984, 2000a;

Bernhardt & Kamil, 1985; Alderson & Banerjee 2002). Si las habilidades de lectura del primer idioma se transfieren al segundo, entonces tendremos que enseñar a leer en el primer idioma con precisión y corrección antes de esperar que lo hagan en un segundo idioma. Sin embargo, la tendencia general en este momento es la de aceptar la hipótesis de Clarke (1979), llamada "hipótesis del cortacircuito" que afirma que las destrezas de lectura de un primer idioma solamente se pueden transferir al segundo a partir de que se haya adquirido un nivel umbral de competencia en ese idioma. El problema ahora es decidir cual es el nivel umbral de competencia, ya que, como hemos dicho anteriormente, la comprensión lectora depende de muchos factores.

Según Alderson & Banerjee, el nivel umbral de competencia a partir del cual se pueden transferir las destrezas de lectura de un primer idioma al segundo dependerá de la dificultad del texto y de la familiaridad del tema del mismo:

> A reader's first language reading skills may "transfer" at a lower level of foreign language proficiency, on a text on a familiar topic, written in easy language, with a clear structure, than they would on a less familiar topic, with much less clearly structure organisation, with difficult language. (Alderson y Banerjee, 2002: 84)

La implicación para la evaluación de la comprensión lectora en un segundo idioma y la interpretación de los resultados es que un mal resultado en la actuación de un candidato se debe probablemente al insuficiente conocimiento de la lengua, más que a un déficit de habilidades de lectura. Una de las metas de la investigación sobre la evaluación es explorar que es lo que hace que un test, una tarea o un ítem sea difícil. De acuerdo con Alderson (2000a), los candidatos que responden a un test lo hacen interaccionando con la tarea de forma variada y compleja, y los expertos que juzgan los tests no están en situación de saber cómo los candidatos van a responder a cada ítem o de hacer generalizaciones sobre qué destreza de la comprensión lectora está midiendo cada ejercicio, ya que no tienen la misma competencia lingüística que los examinandos

3.3.6. El género y la estructura del texto

El conocimiento de cómo se organizan los textos y como se señala la información o los cambios de contenido facilita la lectura. La naturaleza del texto también tiene un claro efecto sobre la dificultad del test. Perkins (1992) al estudiar el efecto de la estructura del pasaje sobre los resultados del test, llegó a la conclusión de que las cuestiones derivadas de frases donde la información conocida precede a la nueva son más fáciles que las cuestiones que se derivan de frases con diferente estructura. Sin embargo, Salenger-Meyer (1991) afirma que las cosas no son tan simples, ya que la familiaridad de los estudiantes con los temas del texto afecta a los resultados más que la estructura del mismo. Cuando los temas no son familiares a los estudiantes, los cambios de estructura afectan sólo a los que poseen un nivel bajo, no a los que tienen un nivel alto. Si el tema no es familiar a los estudiantes, entonces el hecho de que el texto contenga estructuras de nivel no afecta a ninguno de ellos.

Esto nos lleva a pensar que la estructura del texto es un factor de dificultad del mismo, pero que no se puede estudiar aisladamente, ya que se ve afectado por otras características del texto y de los lectores. Perkins & Brutten (1988) estudiaron la relación que existe entre diferentes niveles de comprensión lectora, la estructura de los textos y el conocimiento de los alumnos. Lo que descubrieron fue que los alumnos de nivel alto podían comprender cuestiones cuya fuente de información estaba implícita, mientras que los de nivel más bajo no podían, y que tanto los alumnos con altas habilidades de comprensión lectora como los que no tenían esas habilidades mostraron competencia con las estructuras lingüísticas que relacionaban partes del texto, independientemente de su competencia de la lengua. Esto vuelve a plantear el tema de la relación que existe entre la habilidad de la comprensión lectora y la competencia lingüística.

Otro factor que puede ayudar a la comprensión lectora de los candidatos es el saber encontrar las estructuras léxicas del texto. De acuerdo con Hoey (1991, 2001), los textos no narrativos de inglés contienen una serie de estructuras léxicas que incluyen sinónimos, polisemia, colocaciones típicas, o ciertos grupos de palabras en el texto que él llama *links*, (cuando una palabra de una frase se repite o se parafrasea en otra), o *bonds* (conexiones que existen entre un par de frases). Según su teoría, las palabras que forman la idea principal del texto son reconocidas cuando estas palabras se repiten o se parafrasean en el texto:

> If a learner has acquired – or is shown – the vocabulary used in the central sentences of the text, then he or she would be able to track an intelligible path through unedited authentic text.
> (Hoey, 1991: 231)

De acuerdo con su investigación, Hoey (1991, 2001) afirma que independientemente de la limitación de competencia de la lengua, los que aprenden un idioma extranjero pueden encontrar hasta cierto punto estas estructuras léxicas. Esto tiene importantes implicaciones en la investigación sobre cómo alumnos de ESL o de EFL usan la información contextual en la comprensión lectora.

Podemos terminar subrayando la importancia que tiene la elección del texto para comprender la naturaleza de la comprensión lectora y para diseñar tareas apropiadas que den origen a distintos motivos para la lectura de ese texto. También hemos visto la importancia que tiene el tipo de texto que usemos y el tema del mismo para determinar la validez del contenido.

3.3.7. El vocabulario

El aprendizaje de vocabulario se ha asociado durante mucho tiempo con la lectura. Sin embargo, esta relación no es simple. Laufer (1997b) señala tres

problemas de léxico que pueden impedir seriamente la comprensión lectora en un segundo idioma:

1. Insuficiente conocimiento de vocabulario.
2. Interpretación errónea de algunas palabras de transparencia engañosa, por ejemplo, falsos amigos o expresiones idiomáticas.
3. La falta de habilidad para adivinar correctamente palabras desconocidas.

Laufer afirma que el factor más importante para leer correctamente es el número de palabras que un alumno conoce. Para entender una lectura no especializada se necesitan unas 3.000 familias de palabras que equivalen a 5.000 palabras o unidades léxicas y que cubren alrededor del 95% de un texto. Por debajo de ese umbral las estrategias de lectura no son efectivas.

El adivinar los significados de las palabras usando claves contextuales tampoco es fácil, ya que a veces:

- Esas claves no existen.
- No existe familiaridad con las palabras donde las claves están situadas.
- Las pistas facilitadas son parciales o inducen a error.
- Puede haber incompatibilidad entre el esquema del lector y el contenido del texto.

3.3.7.1. El Vocabulario y su importancia en el nivel de competencia de una lengua

En una situación comunicativa la mayor carga de información recae sobre el léxico. Esta es una de las razones por las que el vocabulario ha pasado de

ser un tema completamente olvidado en la adquisición de un segundo idioma a una posición de importancia, al convertirse los temas léxicos en temas centrales para los lingüistas teóricos. Coady y Huckin (1997), Arnaud y Béjoint (1992), Singleton (1999), Read (2000), Schmitt (2000) y Nation (2001).

Muchos autores consideran que el léxico juega un papel muy importante en la estructura de un texto, especialmente los verbos los cuales determinan muchas veces la estructura gramatical de la frase. Sinclair, y Hoey (2005) proponen nuevas teorías radicales de la lengua para reemplazar a nuestras concepciones tradicionales de la gramática. En lugar de considerar la elección de vocabulario limitada por los espacios disponibles y fijados por la gramática, consideran que el léxico está sistemáticamente estructurado por repetidas estructuras del uso de la lengua.

> By far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of those common patterns. Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patters that make up a text. This is totally obscured by the procedures of conventional grammar. (Sinclair, 1991: 108)

Aunque utilizan diferentes términos, Richards (1976), Nation (1990), Bogaards (2000), y Read (2000) están de acuerdo en que conocer una palabra es conocer sus significados, su morfología, su función, su sintaxis o gramática, sus colocaciones típicas, las palabras con las que puede ir asociada, su pronunciación o los registros y tipos de discurso en los que su uso resulta apropiado. Nation (1990) también considera que conocer la frecuencia de una palabra forma también parte del conocimiento de la palabra, haciéndose eco así de la frase de Richards que dice que: “Knowing a word means knowing the degree of probability of encountering that word in speech or print” (Richards, 1976: 86).

Puesto que el conocimiento de léxico presenta tantos aspectos diferentes, no hay una única forma válida de medirlo. También habría que

definir lo que significa “conocimiento léxico” antes de decidir que técnicas se deben utilizar para evaluarlo. La mayoría de los investigadores están de acuerdo en que el conocimiento léxico es una dimensión continua que consta de varios niveles que empiezan con un vago conocimiento de la forma de una palabra y termina con la habilidad de usar correctamente esa palabra en la producción libre. Es lo que comúnmente se llama conocimiento pasivo y conocimiento activo del vocabulario, y que la mayoría de los autores afirman (Aitchison, 1989; Channell, 1988; y Laufer & Paribakht, 1998) que el primero es más amplio que el segundo.

La relación entre el vocabulario receptivo o pasivo y el productivo o activo no es constante sino que aumenta con la edad puesto que cuando una persona se va haciendo mayor su conocimiento del vocabulario de menor frecuencia aumenta (Nation, 2001). Laufer (1998) demostró que para alumnos de nivel intermedio y avanzado de inglés como lengua extranjera (EFL), el conocimiento de vocabulario activo y pasivo es similar. Esto no ocurría en el caso de estudiantes de inglés como segundo idioma (ESL) los cuales tienen más vocabulario pasivo que activo debido a la gran cantidad de información que reciben, lo cual demuestra que el conocimiento receptivo no se traslada fácilmente a conocimiento activo.

Los estudios dirigidos a evaluar la competencia general de la lengua basándose en el del conocimiento del vocabulario han dado resultados contradictorios. Wolter (2002) utilizó la asociación de palabras como medio de evaluar la competencia en una lengua extranjera. Sus resultados no apoyaban la idea de que la asociación de palabras en un idioma extranjero está claramente ligada con la competencia de los individuos en esa lengua. Sin embargo, Alderson (2007a) afirma que el conocimiento de las palabras y su significado es un componente crucial de la competencia general de la lengua, y que una medida de esa competencia podría ser no sólo cuantas palabras conoce una persona sino cuantas palabras poco frecuentes conoce. Resulta difícil, sin embargo, determinar la frecuencia de una palabra basándonos en nuestros propios juicios por lo que aconseja utilizar las frecuencias “objetivas” de palabras como las que proporciona el BNC (British National Corpus).

Otros autores, lo mismo que Alderson, opinan que el conocimiento del vocabulario puede servir para evaluar la competencia global de un idioma (Meara y Buxton, 1987). De hecho, el test de diagnóstico más conocido en el ámbito europeo como es el DIALANG usa un test de vocabulario para estimar el nivel de "proficiency" de los examinandos (Alderson y Banerjee, 2002).

Read (1997, 2000) y Read y Chapelle (2001) proponen evaluar el vocabulario desde una perspectiva interacionalista. Siguiendo el punto de vista de Bachman (1990), que considera la competencia de la lengua como un conjunto de destrezas comunicativas, afirman que los tests de vocabulario deben ir más allá del conocimiento de listas de palabras descontextualizadas, y que los tests deben tener consecuencias positivas tales como dar a los examinandos el incentivo para que profundicen en su conocimiento de vocabulario o para que desarrollen estrategias de comunicación efectivas. Esto quiere decir que recomiendan los tests de vocabulario que utilicen pruebas integradoras, como puede ser el C-test, en lugar de pruebas de elementos discretos.

3.3.7.2. Importancia del vocabulario en la comprensión lectora

Numerosos estudios muestran que el conocimiento del vocabulario y la comprensión lectora están íntimamente relacionados. Es evidente que los individuos que poseen un vocabulario más amplio consiguen mayor eficacia comunicativa que los que poseen un vocabulario más restringido. Por lo tanto, no es sorprendente que se haya demostrado que medidas de la cantidad de vocabulario conocido correlacionen positivamente con la comprensión lectora.

De acuerdo con Laufer no es posible la comprensión de un texto sin entender el vocabulario.

> No text comprehension is possible, either in one's native language or in a foreign language, without understanding the text vocabulary. …. it has been consistently demonstrated that reading comprehension is strongly related to vocabulary knowledge, more strongly than to other components of reading. (Laufer, 1997b: 20)

Según Alderson (1984) el leer en un segundo idioma presenta un problema de lectura y un problema de lengua. También afirma que hay que alcanzar un umbral de competencia en una lengua para que las habilidades de lectura del primer idioma puedan ser transferidas al segundo. Al interpretar los textos los estudiantes confían primero en el significado de las palabras, después en su conocimiento del tema y por último en la sintaxis, por lo que podríamos decir que la naturaleza del umbral de competencia para poder entender un texto es principalmente léxica (Laufer, 1992).

Freebody y Anderson (1983) afirman que se ha demostrado consistentemente que la dificultad del vocabulario afecta a la comprensión de los lectores de todas las lenguas. Quian (1999, 2002) y (Laufer, 1992) también llegaron a la conclusión de que el vocabulario contribuye de manera única en la predicción de la comprensión lectora. En la investigación de Quian (2002), las correlaciones de los resultados del test de vocabulario explican cerca del 60% de la varianza en los resultados de la comprensión lectora. Quian y Schedl (2004) llegaron a la misma conclusión por lo que afirman que los tests de vocabulario son una forma valiosa de añadir variedad a la batería de tests de comprensión lectora y un modo práctico de medir la comprensión de los alumnos.

Alderson (2000a) afirma que la dificultad del vocabulario en un texto de lectura hace que éste sea difícil de comprender y no se puede compensar con el conocimiento del tema. Según él, el vocabulario es el mejor predictor de la comprensión de un texto y simplificar la sintaxis no convierte necesariamente el texto en más comprensible. Para entender un texto se utiliza una estrategia conceptual más que sintáctica, lo que implica el procesamiento de palabras

léxicas o de contenido. Se sabe desde hace mucho tiempo que la carga de vocabulario es la forma más efectiva de predecir la dificultad del texto.

> Tests of vocabulary are highly predictive of performance on tests of reading comprehension. In studies of readability, most indices of vocabulary difficulty account for about 80% of the predictive variance. In short, vocabulary plays a very important role in reading tests.
> …. clearly vocabulary is important to text comprehension, and thus to test performance. (Alderson, 2000a: 99)

Carnine et al. (1984) determinaron que era más fácil inferir el significado de palabras desconocidas cuando se presentaban en contexto y cuando la información contextual (sinónimos, antónimos, etc.) estaba más cerca de la palabra desconocida.

Read (2000) afirma que la idea de vocabulario en contexto es un concepto crucial. Las palabras no tienen significado cuando están aisladas sino que lo tienen en relación con el resto de las palabras con las que aparecen. El contexto puede cambiar radicalmente el significado de las palabras, convirtiendo palabras familiares en opacas y palabras no familiares en transparentes. El significado de las palabras, por lo tanto, no viene dado sino que tiene que ser negociado. No es sorprendente por lo tanto que exista una correlación alta entre los tests de comprensión lectora y los de vocabulario en los cuales los ítems de vocabulario se presentan en contexto (Schmitt, 1999).

Los resultados obtenidos en el *cloze* o en el C-test se basan no solamente en el conocimiento de palabras individuales sino en la destreza de deducción o inferencia. Existe mucha relación entre recuperar un término léxico de un C-test y adivinar el significado de las palabras desconocidas en un texto de comprensión lectora (Koda, 1997 y Read, 2000).

Podemos resumir este apartado diciendo que no hay duda de la importancia del vocabulario para la comprensión lectora y del valor de la lectura como medio de incrementar el vocabulario (Nation y Coady, 1988).

3.3.7.3. Factores que determinan la dificultad del aprendizaje y la recuperación de las palabras

Cuando memorizamos o intentamos entender una nueva palabra en una lengua extranjera usamos nuestra experiencia y conocimientos previos, que son los que han determinado las categorías y conceptos de significado en nuestra cultura y en nuestra lengua, como puntos de referencia. Es una ley general y básica de todo aprendizaje el asociar elementos y estructuras nuevas con los que ya están almacenados en nuestra memoria.

Para predecir la legibilidad de un texto se sabe que el factor más consistente y significativo es la dificultad del vocabulario. Esto es así en mucha mayor medida que la estructura de la frase (Nation y Coady, 1988). La dificultad del vocabulario se determina de varias formas. Algunos de los factores que influyen son:

1. *La frecuencia de las palabras*. Alderson (20007a) y Aitchison (1989) consideran que la frecuencia de las palabras es crucial para la comprensión de un texto y es evidente que para comprender minimamente un texto se necesitan conocer las palabras que tienen una frecuencia muy alta en la lengua. Diferentes partes del discurso no están representadas igualmente en los distintos niveles de frecuencia, y se observa que suele haber más nombres entre las palabras de menor frecuencia (Nation, 2001).

   Cuanto más frecuente sea una palabra más fácilmente se encontrará en el discurso hablado o escrito, más necesidad

habrá de utilizarla en la comunicación y más fácil será memorizarla. Es lo que se ha llamado “incidental learning” o adquisición de vocabulario de forma inconsciente (Read, 2000 y Ellis, 1997). Tenemos que recordar, sin embargo, que las personas nativas no aprenden las palabras de su lengua materna siguiendo el orden de frecuencia de las mismas. Las palabras funcionales que forman una clase cerrada (preposiciones, conjunciones, artículos, pronombres, auxiliares, etc.) son muy frecuentes pero se adquieren bastante tarde por los niños que están aprendiendo su primera lengua. Por otra parte, una persona que aprende un segundo idioma empieza a utilizar estas palabras muy pronto. Esto es debido a que el desarrollo conceptual determina la adquisición de las palabras y las personas que están aprendiendo un segundo idioma ya tienen la noción de preposición, pronombre y otros tipos de palabras funcionales y por lo tanto usarán este conocimiento cuando tengan que aprender otros idiomas.

2. *La clase de palabras a la que pertenece el término léxico*. Los resultados de varios estudios de la lengua sugieren que ciertos términos léxicos son más fáciles de aprender que otros. Schmitt (2000: 60) afirma que la clase de palabras está ciertamente implicada en el aprendizaje y el almacenamiento del vocabulario. Ellis y Beaton (1993) hallaron que los nombres son los más fáciles de aprender seguidos de los adjetivos y que los verbos y los adverbios son los más difíciles. Esto es debido a que la adquisición de sistemas léxicos implica la habilidad de desarrollar conceptos (Wode et al., 1992: 53), de ahí que las palabras que se aprendan, las que se almacenen, las que se recuperen y las que se activen más fácilmente sean las que crean una representación visual más clara en nuestra mente (Neuner (1992: 164).

Es muy frecuente que aunque los estudiantes adquieran el contenido semántico de algunas palabras cometan errores al confundir la clase a la que pertenecen. Así vemos que a veces el nombre se sustituye por el adjetivo, por el verbo, etc. (Odlin y Natalico, 1982).

3. *La longitud de la palabra*. Algunos autores, por ejemplo Nation y Coady (1988), opinan que las palabras largas son más difíciles de aprender que las cortas. Sin embargo, Laufer (1997a) asegura que no existen resultados empíricos que demuestren esta afirmación. Por el contrario, piensa que muchas palabras largas son transparentes morfológicamente, es decir, que se componen de varios morfemas familiares, ej. mismanagement, por lo que no hay ninguna razón plausible por la que una palabra larga deba presentar ninguna dificultad de entendimiento o memorización. Lo importante a la hora de aprender una palabra es la frecuencia con la que un estudiante se encuentra esa palabra, no si es larga o corta.

4. *Las flexiones y derivaciones.* La falta de regularidad en los plurales o el género de los nombres y en la combinación de morfemas para crear significados añaden complejidad a las palabras y las hacen más difíciles de aprender.

5. *Similitud de las formas léxicas.* Las palabras que se parecen en la pronunciación o en la ortografía suelen ser confundidas por los estudiantes.

6. *La abstracción*. Generalmente se asume que las palabras abstractas son más difíciles que las concretas, ya que las primeras son más complejas que las últimas. Laufer (1997a), sin embargo, asegura que esto es cierto en el caso de la adquisición de la primera lengua cuando el desarrollo léxico y el cognitivo van de la mano. Los estudiantes de una segunda

lengua ya han desarrollado los conceptos abstractos por lo que una palabra abstracta en sí misma no tiene por qué ser más difícil de entender y recordar que una palabra concreta

7. *La especificidad de la palabra.* Las palabras generales y neutras, que pueden ser usadas en una gran variedad de contextos y registros, son menos problemáticas para la producción que las palabras que están restringidas para ciertas áreas o registros específicos.

8. *La lengua materna.* La lengua materna influye en la forma en que se aprende el vocabulario de una segunda lengua, en como se recupera para su uso y en el modo en el que los estudiantes compensan su falta de conocimiento intentando construir unidades léxicas complejas (Swan, 1997). Nesselhauf (2003) llegó también a esta conclusión cuando demostró la gran proporción de errores que los alumnos cometían debido a la influencia de su lengua materna.

9. *Significados múltiples* Existen significados que pueden estar representados por varias palabras y palabras que pueden tener varios significados. En este último caso estamos hablando de polisemia y homonimia que dependen de si los significados están relacionados o no. Los significados de la palabra *neck* tanto si nos referimos a la parte del cuerpo de una persona, al de una botella o al de una camisa están relacionados. Sin embargo, el significado de la palabra *bank* puede referirse a la institución financiera o a la orilla de un río y no estarían relacionados. Lyons (1981: 148) afirma que no debemos preocuparnos por esta distinción, ya que el problema de distinguir entre homonimia y polisemia es, en principio, insoluble.

Lo cierto es que los estudiantes tienen verdaderos problemas con estas palabras, especialmente cuando la polisemia y la homonimia de su primera lengua son diferentes a las del segundo idioma y ellos transfieren las palabras de la L1 a la L2 automáticamente sin tener en cuenta el diferente uso y significado de las palabras

10. *Las expresiones idiomáticas* son mucho más difíciles de aprender y entender que las expresiones no idiomáticas equivalentes, incluso cuando estas expresiones sean similares en la primera y en la segunda lengua.

11. *La carga afectiva que la palabra tenga para el estudiante*, es decir, la relación personal con el contenido de la nueva palabra. Aquí tenemos que tener en cuenta la influencia de la lengua materna en el aprendizaje de la segunda lengua que como hemos visto es muy importante.

12. *La pronunciación*. Las palabras que son difíciles de pronunciar se aprenden más despacio que las que no presentan mayor dificultad en la pronunciación (Ellis y Beaton, 1993). Esto ocurre especialmente entre estudiantes adultos que tienden a evitar pronunciar las palabras fonéticamente difíciles. Una de las consecuencias de la falta de uso de estas palabras hace que su aprendizaje sea más difícil.

13. La *ortografía*. Los estudiantes frecuentemente confunden palabras de la segunda lengua que se pronuncian o se escriben de forma parecida (Laufer, 1979a). El procesar la ortografía de forma ineficiente conduce no sólo a un bajo nivel de recuperación de palabras sino a una mala comprensión del idioma. Se cree que el conocimiento de ortografía juega un papel principal en la comprensión lectora de un segundo idioma, principalmente en el procesamiento del léxico (Koda,

1997). En la mayoría de palabras inglesas no existe correspondencia entre la fonología y la ortografía, por lo que son candidatas para que se escriban o se pronuncien de forma errónea.

14. *Las colocaciones típicas de las palabras.* Un alto porcentaje de todos los errores que comete un estudiante de L2 corresponde a errores de colocaciones típicas, especialmente las formadas por verbos seguidos de nombre, ej. “to revoke a licence”. Ello es debido a que el estudiante transfiere los significados de su primera lengua a la segunda de forma sistemática creando así nuevas expresiones (Biskup, 1992). No es probable que una persona no nativa acierte en la predicción de los términos léxicos que son compatibles para formar una colocación típica. Además, los estudiantes cuando se encuentran con una colocación léxica no la prestan atención, y a veces les pasa inadvertida, ya que no les cuesta ningún esfuerzo entenderla, razón por la que no la memorizan para poder utilizarla más tarde.

15. *La organización de los conceptos en nuestra memoria*. De acuerdo con Neuner (1992: 161-162), las palabras que representan conceptos generales no se organizan en estructuras jerárquicas sino de acuerdo con las distintas relaciones que existen entre los términos léxicos del lenguaje como puede ser la antonimia. Cuando se aprenden dos palabras de significado contrario, ambas palabras son más fáciles de recuperar de nuestra memoria y activarlas para usarlas de forma activa.

3.3.8. El método

Parte de la investigación que se realiza actualmente sobre la evaluación se centra en la influencia del método sobre los tests de comprensión lectora. Las personas que elaboran tests deben tener presente los efectos sistemáticos de los diferentes formatos de examen para evitarlos, o al menos tenerlos en cuenta a la hora de evaluar los resultados o hacer generalizaciones. Algunos investigadores continúan analizando métodos tradicionales, tales como los de elección múltiple, preguntas cortas, rellenar huecos y C-tests. Otros, como Wolf (1993) afirman que el test de "Recuerdo Inmediato" solamente puede evaluar la recuperación de detalles de bajo nivel. Wolf concluye que la habilidad de los estudiantes para demostrar su comprensión depende de la tarea y del lenguaje de las preguntas del test. También afirma que el seleccionar una respuesta (elección múltiple) o el construirla (*cloze, short answers*) mide diferentes habilidades. También sugiere que el hacer las preguntas en el primer idioma puede ser más apropiado que en el idioma que están estudiando para medir la comprensión en lugar de la producción.

Sin embargo, Alderson & Urquarth (1984) comparan dos métodos de evaluación de la comprensión lectora, uno de rellenar huecos y otro de respuestas cortas, hallando que el efecto del método sobre los resultados es considerable. Deville & Chalhoub-Deville (1993) advierte sobre el aumento de los métodos de recordar lo que se ha leído, diciendo que solamente cuando las puntuaciones de estos métodos se sometan al análisis de fiabilidad se podrán considerar como alternativa a otros métodos de comprensión.

Riley & Lee (1996) comparan los métodos de recuerdo y resumen, y concluyen que hay bastantes diferencias cualitativas entre los dos métodos. Los resúmenes contienen más ideas principales que los métodos de recuerdo y estos últimos contienen más porcentaje de detalles que de ideas principales.

La traducción también ha sido investigada como un método para medir la habilidad de la comprensión lectora (Buck, 1992), mostrando

sorprendentemente buenos índices de validez, aunque no se ha seguido investigando sobre este método.

Si la finalidad de los tests es proporcionar una medida precisa sobre la destreza de los que aprenden un idioma, los que diseñan los tests tendrán que minimizar la influencia de los factores que intervienen, tales como la organización del texto y el formato de los tests. Estos dos factores han sido estudiados por Kobayashy (2002) llegando a la conclusión de que ambos tienen una gran influencia sobre los resultados de los candidatos.

Cuando los textos estaban claramente estructurados, los alumnos de mayor nivel obtenían mejores resultados en el resumen y en las preguntas abiertas. Sin embargo, la estructura del texto apenas influía sobre los resultados que tenían un nivel de lengua más bajo. Esto sugiere que los textos bien estructurados ayudan a diferenciar entre alumnos con diferente nivel de competencia. Luego si queremos aumentar el nivel de validez habrá que prestar atención a estos factores.

Chen (2004) mostró su desacuerdo con Kobayashy (2002) argumentando que éste último había utilizado para su estudio diferentes pasajes de lectura, por lo que lógicamente no se podía afirmar si las diferencias observadas en la actuación de los candidatos se debía a la organización del texto o a los diferentes niveles de dificultad de los mismos, debido a los cambios de contenido o a como ese contenido estaba expresado. Como no existe garantía de que todos los pasajes utilizados son comparables, no existen bases sólidas para llegar a la conclusión de que los alumnos con un nivel de competencia bajo no se benefician de los textos con una buena organización y estructura clara.

Como parece que diferentes métodos son apropiados para medir distintos aspectos de la comprensión lectora, la investigación tendrá que continuar tanto revisando los métodos antiguos como introduciendo otros nuevos.

3.3.9. La longitud del texto

La longitud del texto dependerá del propósito de la lectura, pero tenemos que ser conscientes de que los procesos cognitivos serán diferentes en textos de diferente longitud. Por lo tanto, debemos tener en cuenta que cuanto más largo sea un texto, más difícil será procesarlo, ya que se requerirá mayor conocimiento de la lengua:

> In general the longer the text candidates are presented with, the greater the language knowledge that might be required to process it. If short texts are not making the demands on these resources that will occur in normal cognitive processing, theory-based validity is compromised. (Weir, 2005a: 74)

La habilidad para identificar la idea principal en textos largos es cualitativamente diferente a la habilidad para identificar la idea principal en textos cortos. Los exámenes para fines específicos de IELTS utilizan textos largos argumentando que son más parecidos a los que los alumnos se van a encontrar en sus estudios, es decir, que utilizan el argumento de la autenticidad. En cambio los exámenes de TOEFL utilizan varios pasajes cortos argumentando que así pueden cubrir una variedad de temas reduciendo la potencial parcialidad de una gama restringida de temas. Lo lógico sería llegar a un compromiso donde se pueda maximizar la autenticidad y minimizar la parcialidad del contenido.

3.3.10. La presencia del texto mientras se contesta a las preguntas

En algunos exámenes, especialmente en los que se programan por ordenador, no se permite volver a leer el texto cuando se contestan las preguntas. Johnston (1984) constató que la disponibilidad del texto cuando los

candidatos contestaban a las preguntas afectaba positiva o negativamente a los resultados dependiendo del tipo de preguntas.

Si se está interesado en saber si los candidatos pueden contestar preguntas relativamente explícitas, orientadas lingüísticamente, o periféricas a la idea central del tema, se debería permitir que los examinandos puedan consultar el texto. Si, por el contrario, se está interesado en saber si los candidatos han entendido la idea principal, entonces el texto debería ser retirado, ya que según este estudio los resultados mejoraban cuando los lectores no podían consultar el texto de nuevo. Sin embargo, hay que tener presente que al retirar el texto aumenta el papel que la memoria juega en el proceso de responder aunque no en el proceso de comprensión.

Resumiendo, podemos decir que la validez de un test se mejora normalmente si se aumenta el número de textos, de tareas, de temas, de preguntas y de métodos, ya que de este modo se reduciría la influencia que los factores que acabamos de describir tienen sobre los resultados de los tests de comprensión lectora. Sin embargo, esto entra en conflicto con el tema de lo práctico.

3.4. Estrategias y destrezas del uso del idioma en la lectura

El proceso de lectura implica la utilización de diferentes estrategias y destrezas del uso del idioma. Las primeras corresponden a operaciones mentales o procesos que los que están aprendiendo un idioma seleccionan conscientemente cuando realizan una tarea relacionada con la lengua. Las estrategias del uso del idioma, que varían de acuerdo con el contexto, también constituyen estrategias de examen cuando se aplican a tareas de los tests de idiomas (Cohen, 1998: 92).

A la habilidad de los examinandos de utilizar estrategias del idioma se la llama *competencia estratégica* - un componente del uso comunicativo de la

lengua. Los examinandos utilizan estas estrategias para compensar una carencia en algún área del idioma. (Bachman, 1990). En la competencia estratégica Bachman y Palmer (1996: 70-75) proponen un marco donde incluyen un *componente de evaluación / valoración*, un *componente de fijación de objetivos* y un *componente de planificación.*

Existe una relación entre las características de la tarea del test y las estrategias usadas. Exámenes de formato indirecto, por ejemplo, tests de elección múltiple y *cloze* tests, son formatos que no reflejan tareas del mundo real y por lo tanto pueden activar el uso de estrategias únicamente con el motivo de poder hacer frente a los formatos de los tests. Por el contrario, formatos de tests más directos, tales como resúmenes o preguntas abiertas, es probable que no activen estrategias de examen que ocupen el lugar de estrategias genuinas del uso de la lengua.

El propósito de la lectura puede afectar también al tipo de estrategias y destrezas que los lectores utilizan.

> One of the main implications of a purpose for reading is that it guides readers in the selection of their strategies, the range of skills they draw on, and the intensity with which they draw on each skill. (Rupp et al., 2006)

Los conceptos de estrategias y destrezas de lectura están muy solapados y en la literatura se encuentran distintas, y a veces contradictorias, clasificaciones y definiciones. Alderson hace mención a la falta de definición correcta de lo que entendemos por destreza y por estrategia basándose en una reflexión de Grabe (2000: 10-11) en la cual expresa la necesidad de clarificar la terminología que se está usando y se pregunta cuál es la diferencia entre distintas habilidades, estrategias y destrezas:

> What exactly is the difference between a skill and a strategy? between a level of processing and a level of meaning?. How are “inferencing skills” different from “strategies” like ”recognising

> miscomprehension" or "ability to extract and use information, to synthesize information, to infer information"?. Is "the ability to extract and use information" the same strategy (skill) as "the ability to synthesize information?. (Alderson, 2000a: 306)

Weir (2005a: 90) las agrupa y las utiliza indistintamente al describir los tipos de lectura en que se puede dividir la comprensión lectora. Sin embargo, Alderson (2000a), Urquart & Weir (1998) y Clarke (1979) hacen una distinción entre ambas. Para ellos, las destrezas son habilidades lectoras interiorizadas y automáticas que poseen los que estudian un idioma, lo cual facilita la comprensión de lectura tanto en situaciones de examen como en otras situaciones distintas.

Las estrategias, por el contrario, son técnicas conscientes y tácticas empleadas deliberadamente por el lector para llegar a comprender la lectura, por ejemplo, subrayar las palabras clave o utilizar un diccionario. Alderson y Windeatt (1991) investigaron las estrategias de los lectores analizando los protocolos de los alumnos que estaban haciendo un examen de lectura y no encontraron ninguna relación entre las estrategias usadas y el tipo de ítems, ni entre la dificultad de los ítems y la habilidad de los lectores para entender las ideas principales, inferencias y declaraciones directas.

Está demostrado que los lectores ajustan sus estrategias y utilizan el proceso de comprensión que mejor se ajusta a sus propósitos de lectura. Por lo que en un examen los alumnos van a utilizar estrategias que generalmente no utilizarían en otras condiciones de lectura, ya que su propósito es distinto. A estas estrategias se las llama estrategias de examen. Lo que tenemos que evitar, en la medida de lo posible, cuando preparamos una prueba, es que los alumnos no puedan resolver la tarea utilizando únicamente estrategias de examen. Para ello, en la fase de pilotaje, tenemos que cerciorarnos de cual es el constructo de la prueba. A menudo, las estrategias que se utilizan en un examen dependen del formato del test. Por lo tanto, los candidatos a veces confían y se basan en los materiales de preparación de test, considerando que esa es la estrategia más eficaz para realizar con éxito el examen.

3.4.1. Clases de estrategias usadas en la comprensión lectora

Se han realizado numerosas clasificaciones de las estrategias que se utilizan en la comprensión lectora y las más comunes de todas ellas son las clasificaciones binarias. Todas las categorías binarias se parecen en que reflejan estrategias que ayudan a la comprensión de pequeñas unidades de discurso frente a las que ayudan la comprensión de unidades lingüísticas más extensas.

Algunas de las estrategias de lectura, llamadas también estrategias "botton-up", son estrategias locales, basadas en el lenguaje, que se centran principalmente en el significado de las palabras, en la sintaxis de la frase y en los detalles del texto. Son estrategias donde el lector empieza con la palabra impresa, reconoce los estímulos gráficos, los descodifica en sonidos, reconoce las palabras y descodifica los significados. Algunos ejemplos pueden ser:

- Dividir los elementos léxicos en partes más pequeñas.
- Leer buscando detalles específicos.
- Parafrasear los textos originales.
- Establecer correspondencia entre palabras clave y elementos visuales.
- Establecer correspondencia entre palabras clave de vocabulario o frases.
- Usar el conocimiento de estructuras sintácticas o puntuación.
- Usar las claves del contexto local para interpretar una palabra o frase.

A las estrategias de lectura basadas en el conocimiento, que son globales, que se centran principalmente en la organización del discurso, en la idea principal del texto, y en las que los conocimientos y la formación del candidato tienen una gran influencia, se les suele llamar estrategias ”top-down”. Algunos ejemplos de estas estrategias globales pueden ser:

- Reconocer la idea principal, tema o concepto del texto.
- Integrar la información desperdigada por el texto.
- Inferir información.
- Predecir lo que puede pasar en un escenario determinado.
- Reconocer la estructura del texto.

Actualmente, sin embargo, se considera que ninguna de las dos estrategias anteriormente descritas caracterizan el proceso de lectura. Se cree que cada componente del proceso de lectura puede interactuar con cualquier otro componente, sea “botton-up” o “botton-down”. Este nuevo modelo se denomina *modelo interactivo* y considera que los procesos de lectura son *paralelos* en lugar de *consecutivos* (Grabe, 1991), y que el conocimiento lingüístico y el del mundo interaccionan continua y simultáneamente con la aportación visual.

Puesto que la habilidad para utilizar estas estrategias por parte de los estudiantes de ESL procedentes de diferentes orígenes culturales y lingüísticos, no es la misma, es necesario que en las tareas que se propongan para un examen haya un equilibrio entre las dos clases de estrategias (Abbott, 2007).

3.5. Técnicas para evaluar la comprensión lectora

En la literatura de la evaluación se utilizan los términos “técnica de examen”, “método de examen”, y “formato de examen” de forma más o menos sinónima. Otros autores, sobre todo recientemente, utilizan también los términos “tareas” o “tipo de tarea” para referirse al mismo concepto. Nosotros utilizaremos estos términos de forma indistinta o alternativa.

Las técnicas de examen son medios para obtener información con el propósito de hacer inferencias sobre las habilidades de la lengua de un individuo o tomar decisiones sobre el mismo. La muestra que obtengamos debe ser un indicador válido y fiable de la habilidad que nos interesa. Necesitamos que las técnicas que usemos puedan ser corregidas de forma fiable, económicas y que tengan un efecto rebote beneficioso. Todo lo que dijimos para la evaluación es también válido para la comprensión lectora. No hay un test del que se pueda decir que es “el mejor”. No hay un solo método que pueda cumplir la variedad de motivos por los que podemos querer evaluar. Algunas técnicas son más comunes que otras simplemente por razones de conveniencia y eficiencia no por que sean más válidas.

Existen diferentes métodos o técnicas de evaluar la comprensión lectora, debido a que cada una de ellas está relacionada con la obtención de diferentes niveles de entendimiento de los textos o con las distintas habilidades y destrezas que los lectores utilizan para comprender un pasaje. Cuando se lee en la vida real, los lectores responden a los textos de forma muy diferente dependiendo del propósito que tengan. Está aceptado de forma general que es inadecuado medir la comprensión de un texto por un único método. Puesto que varios estudios sugieren que diferentes formatos miden diferentes aspectos de la habilidad de la lengua, se asume que los buenos tests de comprensión lectora deben emplear un número de diferentes técnicas bien sobre un mismo texto o sobre textos diferentes:

> The response format (test method) used for testing language ability may itself affect the student’s score. Since the effects of the

> response format tend to be unpredictable, it can be a potential source of construct-irrelevant variance. The best advice that can be offered is: ensure that more than one response format for testing any ability is used. (Alderson et al., 1995: 44-45)

3.5.1. Técnicas integradoras y técnicas de elementos discretos

Durante los años 70 se creía que la naturaleza de la lengua era tal que no era posible dividirla entre las partes que la componían. Se la llamó la Hipótesis de Competencia Unitaria - 'Unitary Competence Hypothesis' -. Hoy esta hipótesis se ha abandonado, pero el concepto de habilidad global puede ser un concepto útil.

Otros investigadores opinan que la habilidad de comprensión lectora se puede dividir en distintas destrezas que se pueden medir independientemente. Según esta teoría, si quisiéramos saber la habilidad global de un individuo, tendríamos que medir un número de habilidades separadas y después combinar las puntuaciones. Esto apenas seria rentable si quisiéramos usar los resultados para tomar decisiones no muy importantes.

Basándose en estas dos teorías nacieron las técnicas de elementos discretos y las técnicas integradoras. Si quisiéramos medir un aspecto aislado de la habilidad lectora de un individuo utilizaríamos una técnica de elementos discretos. Si por el contrario quisiéramos obtener una idea general de cómo lee un individuo utilizaríamos una prueba integradora.

Algunos opinan que la habilidad de la comprensión lectora no se puede evaluar con una técnica de elementos discretos, ya que al dividirla entre sus componentes distorsionamos inevitablemente su naturaleza. Ellos creen que es más válido un enfoque más global y unitario y afirman que el *cloze* test, que ha sido objeto de considerables estudios, y que se dice que mide la competencia general en un idioma, está recomendado.

3.5.2. El cloze test

El "*cloze* test" fue primero desarrollado por Taylor en 1953 para medir la legibilidad de un texto. La primera vez que se usó este test fue para medir la competencia en la lectura del inglés como lengua nativa. Muchos estudios han establecido correlaciones altas entre la legibilidad medida por fórmulas y la medida por el *cloze*. Taylor afirmó que el *cloze* podía proporcionar una estimación más precisa de la legibilidad, ya que consistía en lectores reales procesando textos.

Más tarde, los partidarios de las técnicas integradoras consideraron al *cloze* como método ideal para medir la habilidad lectora global de un candidato, ya que no se estaba seguro de lo que esta técnica realmente media. Por el contrario, otros argumentaban que precisamente porque no se sabía lo que el *cloze* medía no se podía decir que estaba evaluando una destreza unitaria. Alderson (1980), klein-Braley (1981), Bachman (1985), Oller (1973), Jonz (1991), y Herrera (2001) representan las distintas posiciones de este debate, considerando unos que el *cloze* únicamente mide habilidades lingüísticas de bajo orden y otros que el *cloze* mide estructuras y procesos significativos a nivel de texto.

Las técnicas integradoras se basaban también en la teoría de que la lengua tiene la comunicación como finalidad y por lo tanto la competencia de la lengua debe ser evaluada a través de su uso comunicativo. Entre las pruebas integradoras, el *cloze* es quizás el más representativo y la técnica que durante más tiempo se lleva usando.

Los *clozes* fueron creados como formas rentables de medir la habilidad global en una lengua y también como medios para medir la habilidad de la comprensión lectora. Se crean eliminando de textos elegidos una de cada

cierto número fijo de palabras (generalmente entre 5-12), independientemente de cual sea la clase o la función dentro de la frase de esa palabra. Se pide a los examinandos que restablezcan las palabras que han sido eliminadas. A veces se acepta como correcta una palabra que tenga sentido en el hueco, aunque no sea la palabra exacta que se ha suprimido. Al principio y al final del texto se dejan una o dos frases intactas para proporcionar contexto al pasaje.

A finales de los años 1970 el *cloze* era un test ampliamente aceptado como un método válido y fiable para evaluar la competencia global de los alumnos de ESL y EFL. El *cloze* se utilizaba en exámenes a gran escala debido a que era fácil de construir y de corregir, por lo menos si se aceptaba sólo la palabra exacta. Algunos investigadores lo recomendaron incluso para los exámenes de clase (Heaton, 1975: 122 y Oller, 1979: 348).

La evaluación de un idioma a través del *cloze* tiene sus cimientos en la "psicología Gelstalt", según la cual las propiedades del conjunto influyen sobre la forma en que se perciben las partes. Aplicado al campo de la lingüística podríamos decir que la habilidad global de la lengua es diferente a la suma de sus diferentes habilidades. Oller que en 1979 defendía la indivisibilidad del idioma según la hipótesis de la competencia unitaria que decía que es imposible dividir la habilidad de un idioma entre sus componentes, en 1983 expresaba lo contrario diciendo que "native language ability is such that a single general factor can account for all of its reliable variance".

Otro de los cimientos lingüísticos del *cloze* es la redundancia, una característica normal del lenguaje que nos ayuda a "descodificarlo" más fácilmente. Cuando queremos comunicar algo omitimos lo que creemos que ya se conoce. Es decir, reducimos la cantidad de redundancia o información que no es necesaria para que nos entiendan.

La redundancia reducida requiere que el oyente o el lector tenga que descifrar las partes que se han eliminado del pasaje recurriendo al conocimiento de las características de redundancia de la lengua. La asunción

es que cualquier hablante nativo y adulto tiene la habilidad de recuperar las partes mutiladas de un texto, ya que es competente en la lengua:

> Adult educated native speakers of a language can, in general, make use of the redundancy of their language to restore damaged messages through their knowledge of rules, patterns and idioms of their own language and culture – their competence.
> (Klein-Braley y Raatz, 1998: 2)

De acuerdo con Lee (2008) para rellenar los huecos del c*loze* el lector tiene que razonar y construir significado textual basado en evidencia contextual derivada del texto. El lector construye significado textual usando el conocimiento que posee de las limitaciones gramaticales y semánticas para predecir que puede venir a continuación en una secuencia de lenguaje. Como ya se ha dicho anteriormente, algunas personas que elaboran tests creen que el *cloze* representa un enfoque global y unitario de evaluar la comprensión lectora, mientras que otros son más escépticos y dicen que no podemos estar seguros de lo que mide la técnica del *cloze*.

El *cloze* se empezó a considerar como un procedimiento fiable y válido de construir tests de comprensión lectora:

> .... foreign language testers have tended to regard the *cloze* as an automatically valid procedure which results in universally valid tests of language and reading. (Alderson, 1979a: 220)

De acuerdo con esta asunción extendida, se consideraba que todos los *clozes* eran equivalentes entre sí independientemente del texto en el que se basaran, el sistema de supresión utilizado y el criterio de corrección que se aplicara. Esta suposición se convirtió en la primera causa de crítica del *cloze*.

Alderson (1979a, 1979b, 1980) examinó el efecto de varias variables metodológicas y la validez concurrente del *cloze* con personas nativas y no nativas y llegó a la conclusión de que:

- Los cambios en la frecuencia de supresión de palabras afectaban a la dificultad del *cloze* de forma impredecible.

- Los cambios en la frecuencia de supresión de palabras daban lugar a cambios en las correlaciones de los *clozes*, con lo cual cambiaba la validez de los tests.

- Había personas nativas cuyos resultados eran peores que los de personas no nativas. Esto significaba que el *cloze* no parece medir una habilidad homogénea que es la que poseen los hablantes nativos pero que no poseen los no nativos.

- Existen dudas sobre si el *cloze* puede medir o no habilidades lingüísticas superiores.

Por consiguiente, Alderson demostró que en términos de dificultad y validez concurrente, los *clozes* no se pueden considerar equivalentes, independientemente del texto, la frecuencia de supresión de palabras y el método de corrección utilizado. Consecuentemente, el *cloze* no puede ser considerado una forma fácil de construir tests válidos de modo automático.

3.5.2.1. Problemas que plantean los *cloze* tests

Los *clozes* tienen la ventaja de que son relativamente fáciles de preparar, administrar y corregir. Los coeficientes de validez y consistencia externa son altos. Sin embargo, no son la panacea que se creía que eran. Algunos de los problemas que presentan los *clozes* (Alderson 1979a, 1979b, 1980 y Klein-Braley, 1981) son los siguientes:

a) La elección de la primera palabra seleccionada puede tener un efecto sobre la validez del test, ya que una vez que se suprime la primera palabra, el resto de supresiones es automática.

b) Se pueden elaborar *clozes* completamente diferentes con un mismo texto. Según la teoría de elaboración de *clozes* el punto donde se empiece la supresión de palabras se debe elegir al azar, con lo cual dependiendo de donde empecemos tendremos unas palabras suprimidas u otras, lo cual dará lugar a diferentes *clozes*. La investigación ha demostrado que cinco versiones de un mismo *cloze* dan lugar a resultados del test significativamente diferentes.

c) Muchos ítems no miden la sensibilidad del discurso más allá de la frase por lo que el nivel de procesamiento del texto es a menudo limitado.

d) Puesto que la persona que elabora el test no tiene control sobre las palabras que se pueden suprimir, tampoco tiene control sobre lo que se está evaluando. Algunas versiones del test pueden tener suprimidas muchas palabras de léxico, lo que puede dar lugar a que sean irrecuperables incluso por personas nativas, mientras que otros *clozes* pueden tener un montón de palabras funcionales suprimidas, lo que origina que el restablecimiento de las mismas por usuarios competentes de la lengua sea bastante fácil.

e) El *cloze clásico* o *cloze de ratio fija* es un *cloze* que se ha elaborado eliminando automáticamente una de cada “n” palabras. Este *cloze* no se puede enmendar fácilmente si cuando se pilota comprobamos que es imposible completar algunos de los blancos. Esto significa que el *cloze* no puede ser pilotado o alterado.

f) El corregir un *cloze* puede resultar difícil y poco fiable, ya que puede haber varias respuestas para un mismo hueco y a veces, los correctores no se ponen de acuerdo sobre qué respuestas se pueden aceptar.

g) Puede ocurrir que solamente unos pocos huecos coincidan con aspectos de la lengua que interesen el evaluador.

h) Mientras que la fiabilidad de un *cloze* determinado puede ser alta porque las palabras de los huecos están correlacionadas entre sí, la validez como una medida global de comprensión lectora puede estar cuestionada si el *cloze* origina que la lectura del texto se realice a nivel de palabras más que a nivel textual. (Cohen, 1998).

i) Personas nativas, inteligentes y con formación muestran habilidades muy diferentes a la hora de predecir las palabras que se han suprimido y, lo que es más importante, alguna de estas personas han obtenido bastantes peores resultados que personas no nativas. Esto nos lleva a la conclusión de que la validez del procedimiento, incluso como medida de habilidad global, está en entredicho.

j) La creación de huecos cada cierto número fijo de palabras origina, con mucha frecuencia, que sea imposible recuperar ciertas palabras.

k) La estrategia que utilizan los candidatos es la de resolver un rompecabezas (lo que le puede ayudar a rellenar los huecos) en lugar de centrarse en la lectura y el entendimiento de las ideas principales del texto,

l) El uso del KR-20 y otras posibles herramientas de medida de la consistencia interna para estimar la fiabilidad, probablemente

no sean permisibles, ya que los huecos pueden no ser independientes entre sí.

m) La dificultad del *cloze* depende del texto.

n) Las palabras funcionales se suelen recuperar más fácilmente que las de contenido, por lo que la dificultad de un *cloze* depende de la proporción de palabras funcionales y palabras de léxico o de contenido que sean suprimidas

Klein-Braley (1997:59-60) señala además otros problemas añadidos:

o) La frecuencia con que se suprimen las palabras en un *cloze* clásico es muy alta. Si se requieren 50 huecos para asegurar un nivel de fiabilidad razonable, entonces necesitaríamos textos extremadamente largos.

p) Si se usa solamente un texto, lo cual es la práctica habitual en los exámenes, entonces no podemos asumir que esto sea una muestra representativa de la lengua. Además, puede que la prueba esté sesgada como resultado del contenido del texto.

q) Los elementos del *cloze* son lógicamente textualmente interdependientes. Por lo tanto los coeficientes de fiabilidad son teóricamente poco sólidos, ya que este enfoque estadístico supone la independencia de los elementos. Esto fue también demostrado por Lee (2004).

r) Si se utiliza el método en el que la palabra recuperada tiene que ser la misma que se había suprimido, entonces los *clozes* resultan muy difíciles incluso para hablantes nativos y competentes. Si por el contrario se aceptan como respuestas correctas las palabras que puedan ser posibles o aceptables,

entonces la corrección se alejaría de la objetividad y el tiempo empleado en la misma sería extremadamente largo.

s) Los coeficientes de fiabilidad y validez para grupos muy homogéneos son muy bajos.

t) Parece intuitivamente razonable que un hablante adulto y nativo obtenga resultados mucho más altos en un test pensado para personas que están aprendiendo esa lengua. Esto no ocurre con los *clozes*.

3.5.2.2. Alternativas como solución a los problemas del *cloze* clásico.

3.5.2.2.1. Cloze de ratio variable

Yamashita (2003) mostró en un estudio, aunque con un número limitado de participantes, que las puntuaciones de los tests de rellenar huecos pueden reflejar la comprensión a nivel de texto y defiende el uso de este procedimiento cuando se desea medir la comprensión global.

Dastjerdi y Talebinezhad (2006) nos dan la alternativa al *cloze* clásico (de ratio fija), en el que se crean huecos cada cierto número de palabras. Ellos sugieren que puede ser sustituido por un tipo de *cloze de ratio variable ("chain-preserving deletion –CPD - cloze*"), en el cual las palabras que afectan a la coherencia y a la cohesión del discurso del texto no son suprimidas. En el *cloze* clásico la cohesión está interrumpida, por lo que no podemos esperar que los examinandos comprendan completamente el texto, ya que muchas relaciones de significado están interrumpidas, lo que origina que el discurso de muchos *clozes* sea ambiguo. También dan contestación a las críticas de Klein-Braley (1997) argumentando que:

a) Si se acepta el procedimiento de CPD, la longitud de los textos se podría controlar dividiendo el test en partes que, aunque coherentes, serían más cortas.

b) Si usamos el procedimiento de ratio variable se puede evitar el sesgo de los ítems, ya que podemos usar textos que aunque largos no sean necesariamente específicos. Además los términos que se suprimen serán ejemplos genuinamente representativos de los elementos del texto.

c) En cuanto a la fiabilidad y la validez, el procedimiento propuesto de ratio variable (CDP) puede, hasta cierto punto, controlar la fiabilidad y la validez, ya que se puede controlar la proporción entre los términos estructurales y los de contenido que se suprimen.

Fotos (1991) ha demostrado en una investigación que los *clozes* que se han elaborado cuidadosamente, tienen el potencial de convertirse en herramientas integradoras de evaluación del inglés como lengua extranjera. Incluso sugiere que se pueden utilizar como sustitutos de la expresión escrita y como parte de un examen de competencia global del inglés. Lee (1996) está de acuerdo con ella sobre que el *cloze* se puede usar como una medida integradora de un examen de competencia de la expresión escrita, pero recomienda una preparación muy cuidadosa de los *clozes*.

Jonz (1990) discrepa de algunas de las críticas de Klein-Braley. Afirma que de acuerdo con su investigación, los *clozes* presentan un alto grado de sensibilidad a las relaciones que existen entre las frases y a la selección léxica, y que la clase de conocimiento de la lengua que se requiere para completar un *cloze* es prácticamente la misma que para completar cualquier otro.

La implicación de estos hallazgos es que el procedimiento para elaborar un *cloze* está lejos de ser errático al seleccionar sus elementos. Este estudio sugiere que, a la hora de evaluar la comprensión de la lengua, el *cloze* produce

tests que generalmente son consistentes en la forma en que miden el conocimiento de la lengua de los candidatos. En un estudio posterior, Jonz (1991) utilizó *clozes* para estudiar los procesos de comprensión de una segunda lengua en usuarios no nativos de la lengua inglesa. El análisis de los datos reveló que los lazos entre las frases son particularmente notables en los procesos de comprensión de los hablantes no nativos de la lengua. Es decir, las puntuaciones obtenidas en un *cloze* reflejan de forma significativa la utilización, por parte de los candidatos, de estructuras lingüísticas a nivel de texto así como procesos de comprensión del discurso.

3.5.2.2.2. Cloze natural

Brown (1993b: 93) quiso estudiar el comportamiento de un *"**Cloze Natural**"*. Se considera un *cloze* natural aquel que está basado en un pasaje que no ha sido manipulado por el evaluador. Es decir, un *cloze* que se ha elaborado sin que haya influido la intuición o el conocimiento de la persona que lo elabora sobre la dificultad del pasaje, la conveniencia de los temas, etc. Es decir, sin que se hayan tenido en cuenta los criterios que se usan a menudo para seleccionar un pasaje apropiado para un grupo de alumnos en particular. El propósito de este estudio fue explorar las características de los *clozes* cuando sus variables no son manipuladas por el investigador. Es el primer estudio cuyos pasajes se han elegido al azar de una biblioteca. Los resultados indicaron que los *clozes naturales* no están necesariamente bien equilibrados ni tienen por qué ser válidos y fiables. Así pues, parece ser que la intervención de un experto es necesaria para elaborar *clozes* sólidos y fiables.

3.5.2.2.3. Método de supresión de letras

Kokkota (1988) sugiere un *cloze* innovador que está a mitad camino entre el C-test y el *cloze* racional (cuya descripción se verá más adelante),

llamado "*procedimiento o método de supresión de letras*". La propuesta es un procedimiento nuevo de supresión de letras, en lugar de palabras enteras, de acuerdo con los siguientes principios:

- Al principio de la palabra se pueden dejar sin suprimir un número variable de letras dependiendo de la dificultad de la palabra.

- En las lenguas con flexiones se puede dejar sin suprimir también alguna letra al final de la palabra.

- El número de letras que se dejan sin suprimir depende de la dificultad del término y de si se acepta solamente una respuesta correcta.

- Si la palabra es muy fácil se suprimen todas las letras.

- El número máximo de letras que se pueden dejar sin suprimir, si se trata de una palabra difícil, es la mitad de las letras de que consta la palabra. En este caso el ítem seria semejante a los del C-test.

- El número de palabras entre huecos dependerá de la longitud y de la dificultad del texto. Generalmente serán entre 4 y 6 palabras.

Una importante ventaja del método de supresión de letras es que es posible ajustar la dificultad del ítem, previsiblemente aumentando o disminuyendo el número de letras que se suprimen. La dificultad de este test se encuentra a mitad de camino entre la dificultad del *cloze* y la del C-test.

Todos los descubrimientos que se han mencionado se pueden resumir diciendo que las personas que elaboran tests no deben depender exclusivamente de ningún tipo de *cloze* para producir tests fiables y útiles. La

selección de tests se debe realizar de acuerdo con el propósito del test y se debe pilotar con anterioridad. Para evitar alguno de los problemas que presentaban los *clozes* vistos hasta el momento se crearon otras variantes.

3.5.2.2.4. Tests de rellenar huecos

Los *tests de rellenar huecos o (Gap-filling tests)* (Alderson 2000a) a los que también se les llama *clozes racionales (rational cloze tests, Fill-in tests)* (Bensoussan y Ramraz, (1984), o *Selective deletion gap-filling (*Weir*,* 2005a), son una variación del *cloze* tradicional (en el que se suprimen palabras al azar), ya que el que construye el test decide, de acuerdo con unas bases racionales, qué palabras suprime. Se intenta dejar 5 ó 6 palabras entre huecos para no crear una prueba demasiado difícil. Sin embargo, los *clozes* racionales no se pueden considerar, estrictamente hablando, tests basados en el principio de redundancia reducida, ya que en ellos la redundancia no se reduce de forma aleatoria.

El *cloze* racional fue primeramente desarrollado por Bachman (1985) para medir habilidades lingüísticas específicas en la evaluación de la comprensión lectora. Esto le confiere una alta validez de contenido, ya que permite controlar exactamente el aspecto de la competencia de la lengua que se desea medir. Él distinguió distintos tipos de ítems dependiendo de si el contexto que se requería para recuperar la palabra omitida se encontraba dentro de la oración, fuera de la oración, dentro del texto o fuera del texto. En su estudio estaba implícito que la dificultad de un término dependerá del grado de contexto que se requiera para resolverlo.

El método de rellenar huecos está mucho más controlado por el evaluador, ya que las eliminaciones de las palabras han sido especialmente seleccionadas para evaluar aspectos elegidos del lenguaje. Si el evaluador suprime palabras de léxico o de contenido, por ejemplo, es que intenta evaluar la comprensión del sentido general del texto, mientras que si suprime palabras

funcionales o de estructura lo que intenta evaluar es principalmente la coherencia del texto.

El *cloze* racional y el *cloze* clásico presentan los mismos coeficientes de correlación con otros tests de criterios externos y apenas difieren en sus valores de consistencia interna o fiabilidad. La principal ventaja que presenta el test racional respecto al *cloze* clásico es que puede definir más específicamente lo que intenta medir. Es preciso recordar que todos los *clozes* racionales, por ser de ratio variable (*clozes* en los que no hay un número fijo de palabras entre huecos), tienen la ventaja de poder controlar la dificultad del test, puesto que al controlar el número de palabras suprimidas en un texto dado estamos controlando la reducción de la redundancia del texto. Cuanta más separación haya entre las palabras suprimidas más fácil será el test, ya que tendrá menos huecos para una misma longitud de texto, y por lo tanto la reducción de la redundancia será menor. Los tests de este tipo se suelen considerar tests de competencia general más que de lectura, aunque los análisis estadísticos demuestran que el factor de la comprensión lectora es bastante fuerte a la hora de determinar los resultados de los tests.

Yamashita (2003) afirma que los *clozes* de ratio fija contienen muchos más huecos que pueden ser rellenados simplemente usando el conocimiento gramatical en el ámbito de frase o el conocimiento extra-textual que los huecos que requieren la habilidad de usar las limitaciones impuestas por el texto. Parece por lo tanto razonable utilizar *clozes* de ratio variable en lugar de los *clozes* clásicos cuando se quiera medir la comprensión lectora global, teniendo la precaución además de suprimir las palabras que requieran una comprensión a nivel de texto.

El c*loze* racional ha recibido atención limitada en la investigación de L2, considerándolo principalmente como una herramienta de evaluación (Bachman, 1985; Bensousan y Ramraz, 1984; Koda, 2005; Yamashita, 2003). Koda (2005) recomienda elegir palabras léxicas o de contenido como indicador de los diferentes componentes de la habilidad de la comprensión lectora tales como el uso de claves contextuales, sofisticación gramatical, conocimiento del

contenido e integración de la información. Lee (2008) considera que el c*loze* racional puede ser utilizado también para instruir a los alumnos en la lectura, la escritura y el vocabulario. Alderson (2000a) diferencia claramente estos dos tipos de formato llamando a los tests de ratio variable o tests racionales, "gap-filling tests" y reservando el término de "*cloze*" para los *cloze* tests de ratio fija. Él defiende que mientras que los "gap-filling tests" se pueden utilizar como tests de comprensión lectora, los "*clozes* tests" no.

3.5.2.2.5. Cloze de discurso

El método de *cloze* racional en el que los elementos omitidos marcan relaciones entre distintas proposiciones se llama *"cloze de discurso" (discourse cloze)* porque se suprimen únicamente los marcadores de coherencia del texto. En este tipo de *clozes* es también fácil para el que elabora el *cloze* hacer cualquier alteración, si fuera necesario, después de pilotar y analizar el comportamiento de cada elemento del test, y seguir manteniendo el mismo número de ítems, ya que podemos sustituir alguno de los huecos. Otra ventaja es que las respuestas aceptables son limitadas en número, por lo que es posible obtener una alta fiabilidad en la corrección si elaboramos una clave que incluya todas las posibles respuestas que sean válidas o que puedan ser aceptadas. Sin embargo, parece ser que los candidatos, para dar sus respuestas, se limitan a la información contenida en la frase, a pesar de la intención de los que han preparado la prueba

Finalmente, los *clozes* racionales pueden incluir expresiones idiomáticas y otros grupos de palabras que el lector competente reconoce como unidades léxicas completas. Sin embargo, estos *clozes* todavía tienen alguno de los problemas que tienen los *clozes* clásicos.

a) Es una técnica indirecta, lo que significa que normalmente mide una parte limitada (Conocimiento microlingüístico) de lo que puede constituir la competencia de la comprensión lectora.

b) Al ser una prueba indirecta, resulta difícil generalizar y decidir acerca de la habilidad lectora de los candidatos basándonos en las puntuaciones del test.

c) Debido a la limitada cobertura de contenido, y al restringido procesamiento requerido, es improbable que esta prueba sea un indicador suficiente de la habilidad de lectura de los candidatos.

d) La técnica no tiene un efecto rebote positivo sobre la enseñanza, ya que no es en si misma una medida directa sobre el constructo de la comprensión lectora. Es difícil saber la relación que existe entre este test y un proceso normal de lectura.

e) Las estrategias que utilizan los que hacen el test para realizar la tarea parecen cambiar el foco de los estudiantes desde leer y entender las principales ideas del texto hasta tácticas de resolver rompecabezas, las cuales pueden ayudar a rellenar los huecos.

3.5.2.2.6. Cloze con banco de palabras

El *Banked gap-filling test*, también llamado *banked cloze test* o *matching cloze test*, se creó para evitar el problema de la falta de fiabilidad de la corrección que existía en *los clozes* y los en los tests de rellenar huecos. En este test la persona que prepara la prueba proporciona las respuestas, generalmente en orden alfabético, para que los candidatos elijan. De este modo la prueba se convierte en una prueba de lectura más que de competencia general de la lengua. A veces, junto con la lista de respuestas correctas, se añaden varias palabras extra que sirven como distractores. Esto

se hace para evitar que si un candidato elige una palabra para rellenar un hueco equivocadamente no sea penalizado dos veces, y también para no favorecer, especialmente hacia el final de la tarea, la estrategia de adivinar que palabras van en cada hueco.

El *cloze* con distractores es realmente bastante difícil de construir, ya que hay que asegurarse de que una palabra, que se pretendía que fuera un distractor para un hueco, no sea válida para ser usada en otro hueco.

3.5.2.2.7. Cloze de elección múltiple

Una variante tanto del *cloze* como de la prueba de rellenar huecos es proporcionar para cada hueco elecciones múltiples para que los alumnos elijan la respuesta. Las opciones se pueden insertar en el hueco o colocarlas al final del texto identificándolas con el número del hueco a que corresponden. Este modelo de test fue propuesto por Jonz (1976) como una alternativa al *cloze* clásico para reducir el gasto de los exámenes de clasificación. Quería un examen que redujera al mínimo el tiempo de examen y el de corrección y que al mismo tiempo clasificara a los candidatos con suficiente precisión.

Bensoussan y Ramraz (1984) diseñaron un test con cuatro opciones de elección múltiple para cada hueco. Seleccionaron los huecos de acuerdo con los tres niveles de comprensión lectora de un texto, es decir, los niveles lingüístico, pragmático y textual, siguiendo la teoría del análisis del discurso. El criterio básico para elegir los espacios en blanco es que exista suficiente redundancia en el texto para que un lector competente pueda usar las claves o pistas del mismo para rellenar los huecos con una palabra o expresión adecuada. Ellos eligieron palabras individuales o grupos de palabras cortos (no más de tres palabras) tales como términos funcionales (conjunciones, preposiciones, artículos), términos léxicos (sustantivos, adjetivos, verbos, adverbios), y marcadores de cohesión (“on the one hand..... on the other hand”, “either .... or”, “not only….. but”). Se supone que el reconocimiento por

parte del alumno de estos recursos le permite seguir la secuencia de pensamiento y que la falta de reconocimiento impide su comprensión.

Éste es el primer intento de evaluar a “nivel macro” o textual utilizando la técnica del *cloze* con ítems que incluyen grupos de palabras cortos. Es un estudio interesante aunque sus resultados, debido al tamaño restringido de la muestra, no fueron muy significativos.

3.5.3. Test de elección múltiple

Los tests de elección múltiple son un mecanismo muy corriente para evaluar la comprensión de un texto en un segundo idioma. Se requiere que los candidatos seleccionen una respuesta entre una serie de respuestas dadas, de las cuales solamente una es correcta. Las ventajas de estas técnicas son que los evaluadores pueden controlar el rango de posibles respuestas, que la corrección sea perfectamente fiable, que son pruebas rápidas y económicas, y que pueden ser corregidas mecánicamente.

Sin embargo, los tests de elección múltiple también tienen varios inconvenientes. El más significativo es que los examinandos pueden adivinar la respuesta correcta sin entender completamente el pasaje de lectura, por lo que la validez del test es cuestionable. Algunos investigadores afirman que la habilidad de contestar a un test de elección múltiple es diferente de la habilidad de comprensión lectora. Aparte de esto, la elaboración de un test de elección múltiple es muy laboriosa, requiere ciertas habilidades y hay que asegurarse de que si la clave de respuestas da solamente una respuesta correcta es porque solamente hay una respuesta correcta. Hughes (1989: 60-63) menciona los siguientes problemas asociados con las preguntas de elección múltiple:

- La técnica solamente evalúa que se ha reconocido cierta información.

- El elegir al azar o por eliminación las respuestas puede tener un peso considerable y desconocido en la puntuación del test.
- La técnica restringe severamente lo que se puede evaluar.
- Es muy difícil redactar buenas preguntas.
- El efecto rebote puede ser muy dañino.
- Se facilita el que copien.

Según Bachman y Palmer (1996) el propósito de los tests de idiomas es el de permitir hacer inferencias sobre la habilidad que un candidato tiene en una lengua determinada, la cual depende de dos factores, uno es el conocimiento de la lengua y el otro es la competencia estratégica. Esto quiere decir que los candidatos tienen que saber vocabulario, gramática, ortografía y fonética de la lengua, pero también tienen que saber utilizar ese conocimiento de forma efectiva para comunicarse en unas condiciones de tiempo limitado. Una de las mayores críticas de los tests de elección múltiple es que están centrados enteramente en el componente del conocimiento de la lengua olvidándose de la competencia estratégica Read (2000).

El test de elección múltiple es considerado por un gran número de investigadores como una forma de evaluación parcial y limitada, ya que deja sin evaluar parte del constructo: Heaton, 1988; Hughes, 1989; y Weir, 1990. A pesar de ello, los tests de elección múltiple son una técnica muy popular y goza de gran aceptación – validez aparente – entre las instituciones y los candidatos, aunque algunos examinandos la han criticado por ser demasiado artificial, superficial, e inapropiada para evaluar las habilidades de la lengua (Herrera y Martínez, 2002; Herrera et al., 2003; Herrera, 2004).

Rupp et al. (2006) nos demuestran empíricamente que el pedir a los candidatos que respondan a pasajes de un texto con preguntas de elección múltiple induce procesos de respuesta que son totalmente diferentes a los que utilizarían cuando leyeran en contextos que no fueran de examen. También

demuestra que el constructo de la comprensión lectora es específico del examen y viene determinado fundamentalmente por la selección del texto y el diseño de los ítems.

A pesar de las críticas tan poderosas en contra del test de elección múltiple, esta técnica se sigue y se seguirá utilizando mientras existan exámenes a gran escala que tienen que ser corregidos rápidamente. La responsabilidad de elaborar un test que sea válido recae sobre las personas que tienen que construir estos tests que tienen que conseguir que se ajusten a los valores psicométricos apropiados para asegurar que son válidos y fiables.

3.5.4. Técnicas objetivas alternativas

3.5.4.1. Técnicas de correspondencia o de matching

Las técnicas de matching o de correspondencia múltiple son técnicas objetivas donde hay que establecer la correspondencia que existe entre dos series de estímulos, por ejemplo, casar titulares con sus respectivos artículos o párrafos, títulos de libros con extractos de cada libro o una serie de frases con estímulos visuales.

Esta técnica presenta alguno de los problemas de los tests de elección múltiple y los *clozes*. Son difíciles de construir y como hay que proporcionar distractores junto a los títulos (para evitar que cuando todos los títulos excepto uno se hayan adjudicado, exista sólo una elección), tenemos que asegurarnos que no es posible asignar dos títulos al mismo elemento.

3.5.4.2. Técnicas de ordenamiento

En esta tarea los candidatos tienen que ordenar una serie de palabras, párrafos o frases que han sido desordenados. Esta técnica resulta muy atractiva para evaluar la habilidad de detectar cohesión, gramática compleja y sobre todo, organización del texto. Sin embargo, no está exenta de problemas:

a) Son muy difíciles de elaborar satisfactoriamente, ya que se ha demostrado que ordenamientos alternativos son válidos, aunque no fuera el orden original del autor

b) Son también muy difíciles de puntuar cuando los estudiantes sólo consiguen una ordenación parcial de los elementos. Esto es por lo que las tareas se puntúan o bien como completamente correctas o bien como totalmente incorrectas.

Debido a todos los problemas que estas técnicas originan, Alderson opina que puede que no merezca la pena utilizarlas:

> … the amount of effort involved in both constructing and in answering the item may not be considered to be worth it, especially if only one mark is given for the correct version.
> (Alderson et al., 1995: 53).

3.5.4.3. Elementos dicótomos

Esta técnica es muy popular porque parece fácil elaborar ítems con sólo dos alternativas, por ejemplo, sí / no o verdadero / falso. El gran problema de esta técnica es que los estudiantes tienen el 50% de posibilidades de acertar simplemente adivinando la respuesta. Para obtener información acerca de la habilidad de un candidato, es necesario elaborar un número alto de elementos

para así descontar el efecto del azar. Algunas personas de los que elaboran tests reducen la posibilidad de adivinar la respuesta introduciendo una tercera opción como, por ejemplo, “no se dice”. Sin embargo, cuando se intenta evaluar la habilidad para inferir significado, esta opción puede originar bastante confusión tanto a la hora de dar una respuesta como a la hora de corregir la prueba, lo cual puede disminuir su fiabilidad.

3.5.4.4. Técnicas de edición

Esta técnica consiste en pasajes en los que se han introducido errores para que sean identificados por el candidato. Estos errores pueden ser presentados en formato de elección múltiple o pidiendo a los candidatos que identifiquen y corrijan un error por cada renglón del texto. La naturaleza del error determinará si se está evaluando la habilidad de la comprensión lectora u otra habilidad lingüística.

3.5.5. Enfoques integrados alternativos

3.5.5.1. Tests de respuestas cortas

Los tests de respuestas cortas son generalmente los que requieren que los candidatos escriban respuestas cortas en los espacios que se les proporciona. Es una técnica semi-objetiva y puede ser una alternativa a los tests de elección múltiple. Si un candidato acierta la respuesta, podemos asumir que ha entendido el texto, mientras que en las preguntas de elección múltiple los candidatos no justifican la respuesta que han seleccionado y puede ocurrir que hayan elegido una por la eliminación del resto.

Aparte de la lectura detenida, esta técnica puede también evaluar la búsqueda de ideas principales, la lectura superficial para captar la idea fundamental y la lectura rápida para buscar información específica. Evaluar estas estrategias generalmente no es posible con técnicas indirectas tales como las de rellenar huecos. Al elaborar el test debemos asegurarnos que la respuesta se obtenga exclusivamente a través de la información que proporciona el texto y no a través del conocimiento que los candidatos tengan del mundo o gracias a que se puede copiar literalmente del texto.

Sin embargo, esta técnica, que Bachman y Palmer (1996) clasificaron como “limited production response type”, también presenta algunos problemas.

a) Las preguntas que den lugar a respuestas cortas son difíciles de construir, ya que tienen que ser cuidadosamente formuladas, de forma que se puedan predecir todas las respuestas posibles y controlar la extensión de lo que se pueda escribir.

b) Es muy difícil predecir todas las respuestas y evitar ambigüedades en las preguntas.

c) La persona que elabora la prueba tiene que evitar que se pueda dar una respuesta simplemente copiando una frase del texto o gracias al conocimiento que el candidato tenga del mundo.

d) La variabilidad de las respuestas puede originar una falta de fiabilidad en la corrección. Los correctores tendrán que juzgar si las respuestas de los candidatos demuestran que han entendido o no.

e) Hay que dar instrucciones muy concretas a los correctores y, si es posible, proporcionarles una variedad de respuestas alternativas.

f) Es preciso asegurarse de que los correctores no penalizan los errores gramaticales, los de ortografía o los de puntuación. Por lo tanto, es necesario que exista una buena estandarización entre los correctores.

g) Este formato de respuesta implica que los candidatos tienen que dar sus respuestas por escrito, y existe preocupación por si esto interfiriera con la medida del constructo que se desea evaluar.

3.5.5.2. Tests de recuerdo libre o recuerdo inmediato

En este tipo de tests se pide a los estudiantes que lean un texto y que luego escriban todo lo que recuerden del mismo. Se dice que esta técnica proporciona una medida más pura de comprensión lectora, ya que las preguntas no interfieren entre el lector y el texto. También se asegura que proporciona información sobre los procesos que ocurren en la persona que está aprendiendo un idioma nuevo. La descripción de lo que recuerdan tiene que ser en la primera lengua, ya que de lo contrario se convertiría en una prueba de expresión escrita además de comprensión lectora.

Los problemas de este test son:

- Que es difícil de puntuar, lo cual puede originar que la fiabilidad de la corrección no sea buena.

- Que se pueda objetar que éste es un test de memoria más que de comprensión lectora, aunque si la tarea se realiza inmediatamente a la lectura, esto no tiene porqué ser así.

3.5.5.3. Tests de preguntas abiertas

Se dice que los tests de preguntas abiertas evalúan la mayoría de las estrategias y destrezas de lectura, mientras que los *clozes* miden sólo la comprensión local y no reflexionan sobre la comprensión global del lector. Sin embargo, también presentan varios problemas:

a) Estas tareas permiten que los candidatos copien las respuestas directamente del texto, por lo que los correctores no pueden asegurar si el candidato, en realidad, entiende el texto.

b) La fiabilidad de la corrección puede ser baja.

c) Se puede convertir en una prueba de expresión escrita en lugar de una de comprensión lectora.

3.5.5.4. Transferencia de información

Con la finalidad de evitar la contaminación de las puntuaciones las tareas de transferencia de información se pueden usar cuando los candidatos tienen que completar un diagrama, un cuadro o numerar una secuencia de sucesos. Estas tareas a veces se asemejan a actividades de la vida real y son muy usadas en exámenes en los que hay una batería de pruebas y que intentan incluir tareas auténticas.

Los problemas que estas tareas pueden originar son:

a) Puede que los candidatos entiendan el texto pero no tengan completamente claro lo que tienen que hacer en la fase de transferencia.

b) La tarea puede estar cognitiva o culturalmente sesgada. Actualmente, nuestros alumnos pertenecen a distintas culturas, son de diferentes edades, profesiones o nacionalidades. Se puede dar el caso de que la tarea que les mandemos hacer sea una actividad cotidiana para un grupo de personas o para una cultura determinada, pero no para todos los grupos o para todas las culturas.

c) A veces se modifican los textos para que se ajusten más a los requisitos de estas técnicas y otras veces se escriben textos expresamente para ser usados en estos exámenes, por lo que la condición de autenticidad de los textos no se cumple. (Weir 1990).

d) La técnica de transferencia de información añade un elemento de dificultad que no está en el texto.

e) A veces las respuestas requieren un grupo de palabras o frases cortas que tienen que ser corregidas subjetivamente.

Alderson (2000a: 248) afirma que puesto que las personas tienen que realizar en la vida real tareas similares a las propuestas en los tests entonces el sesgo está justificado e incluso es una indicación de validez (”the bias is justified and is, indeed, an indication of validity”). Weir (2005a) y Alderson et al. (1995) consideran que las pruebas de transferencia de información son una variante muy útil a la técnica de respuestas cortas. Las preguntas planteadas en esta técnica cubren la información importante de un texto (información general, principales ideas y detalles importantes) y están relacionadas con los motivos normales por los que la gente lee. Por lo tanto, los candidatos se ven forzados a activar la mayoría de las estrategias y destrezas de la lectura. Esto evidencia que estas dos técnicas se ajustan a la validez basada en la teoría y en el contexto. Otra ventaja de la técnica de respuestas cortas y de la de

transferencia de información es que tanto los textos como las tareas se pueden seleccionar de acuerdo con el nivel de los alumnos.

3.5.5.5. El resumen

La prueba de resumen es una variante del test de recuerdo inmediato. Los examinandos leen un texto y después se les pide que resuman las ideas principales, bien del texto entero o bien de una parte del mismo. En esta prueba se utilizan estrategias genuinas del uso de la lengua en lugar de usar estrategias de examen, y se cree que los candidatos necesitan entender la información principal del texto tanto para separar las ideas relevantes de las irrelevantes como para organizar sus ideas sobre el texto, y así poder realizar la tarea.

Los problemas que puede tener esta técnica son:

a) La corrección de la prueba es subjetiva, ya que el ponerse de acuerdo sobre las ideas principales del texto puede ser casi imposible incluso para lectores expertos.

b) Los candidatos pueden copiar las respuestas directamente del texto.

c) Los candidatos pueden entender el texto pero no son capaces de expresar sus ideas por escrito.

d) Puede que estemos evaluando la habilidad de la expresión escrita en lugar de la comprensión lectora.

Algunos de estos problemas pueden desaparecer si el resumen se puede relacionar con una tarea del mundo real, ya que seria más fácil el poder establecer el grado de adecuación de la respuesta. Otra

solución sería permitir que los candidatos hagan el resumen en su propia lengua en lugar de en la segunda lengua. Sin embargo, en este caso podría surgir la duda de si se está usando la técnica para evaluar la comprensión lectora de la primera lengua. Alderson (2000a) propone que una solución al problema de contaminación de la comprensión lectora con la expresión escrita podría ser el presentar múltiples resúmenes, donde la tarea del candidato sería el seleccionar el mejor resumen de entre todos los proporcionados.

3.5.5.5.1. El resumen de huecos

Para superar las objeciones, antes mencionadas, de la prueba de resumen se ha creado el resumen de huecos. Los candidatos leen el resumen del texto, del cual se han suprimido algunas palabras importantes que tienen que ser restituidas. Para ello tienen que leer y entender las ideas principales del texto. Para tests de comprensión lectora de alumnos de segunda lengua o lengua extranjera, los resúmenes y las respuestas que se piden pueden incluso estar en la primera lengua del candidato. Alderson et al. (1995: 61) sugieren que para evitar el problema de tener varias respuestas alternativas, se puede proporcionar a los candidatos un banco de posibles palabras, como en el "banked gap-filling" que hemos visto anteriormente. El nuevo problema ahora es que estas pruebas son difíciles de construir y hay que pilotarlas muy bien con anterioridad.

3.5.5.6. "*Cloze* elide" tests

Esta técnica fue inventada por Davies en los años 60 no como una medida de comprensión sino como una medida de la velocidad con la que los lectores pueden procesar el texto, y se conocía como "la técnica de la palabra intrusa". A los candidatos se les pedía que detectaran el mayor número de

palabras insertadas posible en un periodo de tiempo limitado. El número de inserciones correctamente identificado menos el número de palabras incorrectamente identificadas daba la medida de la velocidad de lectura. Davies asumía que se requiere cierto grado de entendimiento para identificar las inserciones (Davies, 1975).

En los años 1980 se redescubrió de nuevo el método y se le llamó “cloze-elide technique”. En esta prueba el evaluador inserta palabras en el texto, en lugar de suprimirlas. El examinado obtiene un punto por cada palabra que no pertenece al texto y es correctamente suprimida, y se le quitan puntos por cada palabra suprimida erróneamente. Los puntos en los que se introducen palabras irrelevantes en los pasajes se determinan usando unas tablas con números aleatorios. Se prepara también una tabla con las palabras que se van a añadir al pasaje y que se eligen en el orden en que aparecen en esta tabla y son insertadas en los puntos que determina la tabla de números aleatorios. Si alguna de las palabras que se han introducido tienen sentido en el contexto del pasaje se sustituyen por otras que no lo tengan.

Este test es bastante difícil de elaborar, ya que hay que tener cuidado que las palabras añadidas no den un nuevo sentido al texto. Además el que elabora el test tiene que intuir que clase de comprensión se requiere para identificar la inserción. Aunque este test puede medir habilidades lingüísticas y cognitivas, sin embargo, se considera básicamente como un test de comprensión lectora, (Alderson, 2000a).

3.5.5.7. El C-test

El C-test lo mismo que el *cloze* forma parte de las técnicas de redundancia reducida integradas. Al ser objeto de estudio de esta tesis, este método lo vamos a analizar más detenidamente en el siguiente capítulo.

3.5.6. Autoevaluación

La autoevaluación está cada vez más considerada como una fuente de información sobre las habilidades y los procesos de los alumnos. Ross (1998) demuestra que existen correlaciones de 0.7 o más altas entre la autoevaluación de una habilidad en una lengua extranjera y el test de esa habilidad. El Marco Común de Referencia Europeo ha facilitado las escalas ya conocidas con el nombre de "Can-do" que expresan las funciones de la lengua que un examinando tiene que ser capaz de realizar con éxito en cada uno de los 6 niveles de competencia que se han definido.

La autoevaluación va a ser muy útil en temas de diagnóstico de la competencia de alumno en comprensión lectora. Se prevé un aumento considerable en los trabajos de investigación que relacionen la correspondencia existente entre los resultados de una habilidad medida con un test y los obtenidos con una técnica de autoevaluación.

3.5.7. Técnicas basadas en el ordenador

Inevitablemente tenemos que considerar el papel que hoy en día puede jugar el ordenador en la evaluación de la comprensión lectora. El test de diagnóstico de destrezas se puede hacer fácilmente por ordenador gracias a varios programas como puede ser el DIALANG, que ya hemos comentado. También hemos hablado de las ventajas e inconvenientes que el ordenador puede tener cuando se utiliza para la evaluación de idiomas. El problema que puede existir actualmente es el poder establecer que las variables medidas por ordenador están realmente relacionadas con el uso de las hipotéticas estrategias usadas por el lector al realizar una tarea determinada. Esto es importante para decidir sobre la validez del uso de ordenadores en la evaluación de la comprensión lectora en un segundo idioma

3.6. Conclusiones sobre las técnicas de evaluación de la comprensión lectora

Después de presentar la variedad de métodos o técnicas existentes para evaluar la comprensión lectora y de saber que diferentes métodos son apropiados para medir diferentes aspectos de la comprensión, diremos que actualmente el consenso general entre los investigadores es el de que es esencial utilizar más de un método y más de un texto cuando se intenta medir un constructo como la comprensión lectora, ya que el uso de únicamente una técnica distorsionaría el proceso mismo de lectura.

> Good reading tests are likely to employ a number of different techniques, possibly even on the same text, but certainly across the range of texts tested. This makes good sense since in real- life reading, readers typically respond to texts in a variety of different ways. (Alderson, 2000a: 206)

Actualmente, como hemos visto, resulta problemático alcanzar un consenso sobre lo que constituye la esencia de la comprensión lectora. Los diferentes métodos o técnicas de evaluar la comprensión lectora están basados en diferentes teorías sobre lo que se considera que es el constructo de esta destreza. Sin embargo un área que ha recibido relativamente poca atención por parte de los investigadores es cómo sabemos que una persona ha entendido el texto. Sarig (1989) se enfrenta al problema de significado variable del texto, señalando que diferentes lectores pueden construir diferentes significados de un texto y, sin embargo, acertar la respuesta. Pero esto, no nos ayuda a saber cuando un lector ha interpretado el texto correctamente.

Todos los modelos y técnicas de comprensión lectora descritos anteriormente sugieren que la clave para una comprensión exitosa es la habilidad del lector para extraer y organizar la información de los textos de forma automática, precisa, y eficiente e integrarla con el conocimiento existente de antemano en su cerebro para formar una representación mental coherente

del texto (Esteban, 2005 y Rupp et al., 2006). La desventaja principal de todos los métodos analizados hasta ahora es que no mantienen mucha relación con la forma en que la gente lee en la vida real ni con los motivos por los que leen. Esto puede originar que el test no refleje como los alumnos entenderían el texto en el mundo real.

Puesto que, según hemos visto, los exámenes no pueden reflejar adecuadamente la amplia gama de destrezas que los estudiantes de ESL emplean en la vida real, de acuerdo con Abbott (2007) se debe confiar en los expertos para que determinen que destrezas debemos evaluar y que métodos debemos utilizar para evaluar esas destrezas. Ellos poseen los conocimientos para poder priorizar y seleccionar qué se evalúa, cómo se evalúa y por qué se evalúa.

Resumiendo, como se ha visto en este capítulo la evaluación de la comprensión lectora estará sesgada si solamente usamos una técnica. Diferentes técnicas miden diferentes aspectos de la habilidad de la lectura y puesto que sabemos que cualquier técnica que utilicemos será imperfecta, deberemos siempre tratar de usar múltiples técnicas y métodos cuando evaluemos destrezas de comprensión lectora.

Sin embargo, como ya se ha dicho, todos los métodos anteriormente mencionados para evaluar la habilidad de la comprensión lectora tienen poco que ver con la forma en la que la gente lee textos en el mundo real. Por lo tanto, aún cuando usemos varias técnicas, puede que éstas no reflejen la forma en que los alumnos entenderían el texto en la vida normal, pero lo que sí sabemos ahora, es que el motivo para la lectura y la elección de los textos además de la motivación de los candidatos determinarán el resultado de la comprensión lectora.

Capítulo 4

EL C-TEST COMO PRUEBA DE COMPRENSIÓN LECTORA Y DE COMPETENCIA DE LA LENGUA

4.1. El C- test y los tests de redundancia reducida

De acuerdo con Klein-Brakey y Raatz (1984), Klein-Braley (1985), Klein-Braley (1997) y Katona y Dörney (1993), un C-test es una prueba escrita integradora de la competencia global del idioma basada en el concepto de redundancia reducida. Un C-test consiste en un número variable de textos, generalmente de cuatro a seis textos auténticos, cada uno de ellos completo con un sentido de unidad en sí mismo. En estos textos la primera frase se deja sin alterar. Después se aplica "la regla del dos" que consiste en suprimir la segunda mitad de cada dos palabras empezando por la segunda palabra de la segunda frase. Si una palabra tiene un número impar de letras, se suprime una letra más que la mitad de la palabra. Por ejemplo, "*beach*" se convertiría en "*be*____". Palabras de una sola letra tales como "*I*" no se las tiene en cuenta en el cómputo. Los números y nombres propios se dejan inalterados. Por lo demás, la supresión es enteramente mecánica. El proceso se continúa hasta que tengamos el número de blancos necesarios (20 ó 25 por cada texto para producir un total de 100 elementos o blancos en el C-test completo). El resto del texto se deja inalterado hasta el final.

Los textos se ordenan intuitivamente de acuerdo a su dificultad, colocando el texto más fácil en primer lugar y el más difícil al final. Los tests se corrigen dando un punto a cada palabra correctamente recuperada. Los errores de ortografía se cuentan como tales. La nota final se calcula sumando las de cada texto individual, siendo 100 la nota máxima. La interpretación de la nota puede ser interpretada con referencia a la norma o de acuerdo con un criterio. En términos absolutos, este número se puede relacionar de alguna forma con el nivel absoluto de competencia alcanzado por el examinando.

Los C-tests pertenecen, como ya hemos dicho, a la familia de tests de redundancia reducida, que también incluye a los *clozes*, el dictado, el dictado parcial (consiste en reconstruir las partes de un texto oral que han sido

suprimidas), los *cloze elide tests*, y el "*test de ruido*" de la comprensión auditiva. El concepto de la redundancia deriva de la Teoría de la Información, que dice que un mensaje que es redundante contiene más información de la estrictamente esencial para entenderlo. Esto asegura que si partes del mensaje se pierden o se dañan, bien por ruido o bien por accidente, pueden ser recuperadas a partir de las partes que se han transmitido intactas.

Klein-Braley y Raatz nos recuerdan algún ejemplo como son las fotocopias defectuosas o los anuncios por los altavoces en una estación.

> Redundancy is a feature of natural language since it is quite common for parts of a message to be distorted or missing. Announcements over station loud speakers or copies produced by defective photocopies are obvious examples of damaged communication. (Klein-Braley y Raatz, 1998: 2)

Redundancia es también una característica del conocimiento del mundo: si dos personas comparten la misma formación u orígenes, el mensaje puede ser transmitido mucho más elípticamente que si ese conocimiento común no existiera. El conocimiento cultural es otro tipo de conocimiento que también es relevante para el aprendizaje del idioma. De todo esto, se puede asumir que una persona que está aprendiendo un idioma, y que por definición no posee una competencia completamente desarrollada, será menos capaz de utilizar redundancias para restaurar el mensaje.

En este enfoque de evaluación no se mide la redundancia del texto sino la habilidad del candidato para hacer uso de la redundancia general de la lengua en su conjunto para, de este modo, poder restaurar el texto dañado. Los tests de redundancia reducida pertenecen a los tests posmodernos o psico-lingüísticos de Spolsky. Estos tests fueron diseñados como una reacción al comportamiento tan poco parecido al lenguaje que exigían los tests de elección múltiple. Los tests psico-sociolingüísticos son generalmente técnicas integradoras y están basados en la teoría del lenguaje. Se pretende que sean

modelos de comportamiento lingüístico auténtico, pero no se pretende que sean necesariamente comunicativos.

En 1968 Spolsky et al. dieron a conocer sus investigaciones sobre la evaluación de la lengua usando técnicas que reducían la redundancia disponible en los textos. Spolsky (1973) demostró que personas no nativas, aunque muy competentes en la lengua, obtenían peores resultados que los nativos en tareas de dictado cuando se añadía ruido a la señal. Él afirma que estas diferencias se deben a que los no nativos no tienen experiencia suficiente con la lengua para adivinar las palabras mutiladas.

> The non-native's inability to function with reduced redundancy, evidence that he cannot supply from his knowledge of the language the experience on which to base his guesses as to what is missing. In other words, the key thing missing is the richness of knowledge of probabilities – on all levels, phonological, grammatical, lexical, and semantic –in the language.
> (Spolsky, 1973: 170)

Spolsky llegó a la conclusión de que los tests de redundancia reducida son técnicas que obligan al candidato a actuar de acuerdo a su competencia, por lo que discriminan bien entre nativos y personas muy competentes en la lengua. Spolsky y sus colegas encontraron además buena consistencia interna y considerables correlaciones con otras medidas de competencia global de la lengua.

Para Klein-Braley, el texto que se use para examinar es completamente irrelevante. Lo que los tests de redundancia reducida aspiran a conseguir es obtener una muestra del comportamiento del candidato al azar basándose en un texto elegido también al azar.

> The technique of random sampling is the cornerstone of the theory underlying tests of reduced redundancy: random sampling of the elements of the language through the text forces the examinee to

> exhibit a random sample of linguistic performance in the test.
> (Klein-Braley, 1985: 80)

En cuanto al tema de autenticidad, Klein-Braley cree que no hay tests que puedan ser considerados auténticos. “Lengua auténtica” es la lengua normal que se usa todos los días, y tenemos que aceptar que la lengua normal no se produce para ser evaluada. Nuestra única preocupación debe ser hasta que punto el test de idioma refleja habilidades de la lengua. Como la situación natural de redundancia reducida no se puede trasladar al escenario del examen, tenemos que aceptar que los tests de redundancia reducida de los que disponemos no son sino simulaciones de la realidad.

> The tests of reduced redundancy which are available are therefore all simulations of reality. Attempts to model the real situation under the controlled conditions of the test. They use authentic materials (written texts or recorded utterances), damage them in some way, and present them to the examinee to be restored. The examinee’s processing of the mutilated text enables conclusions to be drawn about his or her level of language proficiency.
> (Klein-Braley, 1997: 48).

Tests como el *cloze* o el C-test, por lo tanto, pueden reclamar relativa autenticidad, ya que el material del test es auténtico y el comportamiento que se pide al candidato se considera como una aproximación al comportamiento lingüístico que se necesita en la vida diaria.

4.2. Características del C-test

El C-test fue concebido como una mejora técnica sobre los *clozes*. En 1982 Raatz y Klein-Braley presentaron el C-test como un intento de retener los aspectos positivos de los *clozes* y la teoría en la que se basaban y remediar sus defectos, mejorando el proceso de muestreo en el desarrollo del test y por lo tanto en la actuación de los sujetos. Raatz y Klein-Braley (1982)

establecieron las condiciones que debía cumplir el nuevo formato, un resumen de las cuales serían las siguientes:

- El nuevo test debe ser mucho más corto, pero al mismo tiempo debe constar de al menos 100 elementos (palabras mutiladas).

- La ratio de supresión y el punto donde se empiezan a suprimir o dañar las palabras debe ser fijo.

- Las palabras afectadas por las supresiones deben ser una muestra representativa de los elementos del texto.

- No se debe favorecer a los examinandos que posean algún conocimiento especial, utilizando textos específicos. Por lo tanto, el nuevo test debe consistir en un número de textos diferentes.

- Solamente la corrección exacta es posible para asegurar la objetividad.

- Los hablantes nativos deben ser capaces virtualmente de obtener puntuaciones máximas (90% o mayores). Si los hablantes nativos no pueden obtener puntuaciones superiores al 90%, entonces el texto no debería usarse para hablantes no nativos.

- El nuevo test tiene que ser fiable, válido y fácil de elaborar.

Los primeros resultados, presentados en 1982, mostraban que se cumplían todas estas condiciones (Raatz y Klein-Braley, 1982). Ellos demostraron que hablantes nativos, adultos y educados, obtenían virtualmente puntuaciones perfectas en los C-tests y que el método de supresión de palabras del C-test producía muestras al azar de las clases de palabras que contenía el texto en cuestión. Investigaciones posteriores (Raatz 1984; Klein-Braley, 1984) confirmaron estos tempranos hallazgos.

4.3. Ventajas del C-test sobre el *cloze* test

Aunque no se ha realizado tanta investigación sobre el C-test como sobre el *cloze*, sin embargo, existen unos pocos estudios sobre los puntos fuertes y débiles del C-test. De éstos, los que se han elaborado siguiendo los principios establecidos anteriormente tienen las siguientes ventajas: resumidas por Klein-Braley (1997: 65). Alderson (2000a: 225), Hughes (1989: 71) y Connelly (1997: 142):

1) Muchos más elementos son posibles con textos mucho más cortos. Para elaborar un *cloze* clásico, que suprima una palabra cada cinco, tendríamos que disponer de un texto de por lo menos 500 palabras de largo para que tuviera 100 elementos. Un C-test que tuviera 5 textos con 20 palabras suprimidas por la mitad sería aproximadamente la mitad de largo.

2) La corrección del C-Test es exacta y objetiva, ya que casi siempre hay solamente una posible solución.

3) La corrección del C-Test es rápida y fácil para el hablante nativo o para el profesor, ya que solamente se tarda un tiempo ligeramente superior al necesario para simplemente leer el texto

4) Los C-tests son muy fáciles para los hablantes nativos. Por otra parte alguien que no entienda el idioma en absoluto, normalmente obtiene una puntuación de cero o cercana a cero.

5) Puesto que una de cada dos palabras está mutilada, la probabilidad de obtener una muestra representativa de todas las clases de palabras que hay en el texto es mucho más alta.

6) Debido a que el C-Test está formado por un número de textos diferentes, la muestra de clases de contenidos es mejor. Los examinandos que tengan un conocimiento especial en ciertas áreas ya no tienen ventaja sustancial sobre el resto de los candidatos.

7) Se dice que esta técnica es más fiable y una medida de comprensión más integral que el *cloze*.

8) Los textos relativamente cortos hacen que el test sea más manejable ofreciendo, sin embargo, una buena consistencia interna.

9) Es posible elaborar un test que tenga una amplia gama de estilos y niveles de habilidad.

10) Las investigaciones parecen indicar que el C-test funciona bien como medida aproximada de la habilidad total en una lengua extranjera.

11) Incluso el material que no ha sido tratado previamente produce tests con coeficientes de fiabilidad y validez satisfactorios.

12) Los C-tests pueden producir resultados que son psicométricamente superiores a los de los *clozes*.

Resumiendo, para los centros y departamentos que tengan que evaluar la competencia global de una lengua de forma efectiva, el C-test

es una elección atractiva por varia razones: tiene un diseño relativamente fácil, no es costoso de construir o pilotar, son cortos, fáciles de administrar, y se pueden corregir rápida y eficientemente por unos pocos correctores.

4.4. Desventajas del C-test sobre el *cloze* test

Todas las técnicas que hemos visto hasta ahora y que se utilizan para evaluar la comprensión lectora tienen ventajas sobre otros métodos pero también presentan algún problema. Lo mismo ocurre con el C-test. Hughes (1989: 71), Alderson (2000a: 225), y Connelly (1997: 142) mencionan los posibles problemas relacionados con el C-test:

1 La tarea tiene una apariencia de rompecabezas.

2 Es más difícil de leer que un pasaje de *cloze*

3 El C-test tiene poca validez aparente con algún grupo de estudiantes y profesores. La apariencia del texto mutilado es inicialmente desalentadora y no se parece en nada que aparezca en el mundo real de la persona que está aprendiendo un idioma. Jafarpur (1995).

4 Muchos lectores encuentran el C-test todavía más irritante que el *cloze*.

5 Hay dudas sobre la validez de su constructo.

6 Existen dudas sobre lo que mide realmente el C-test. (Este problema es el mismo que existe con el *cloze*).

7 Hay alguna evidencia de que el C-test tiende a evaluar el conocimiento de gramática y vocabulario en lugar de unidades de discurso más largas. – aunque esta tendencia fuera verdadera, esto no perjudicaría al test a los ojos de algunos evaluadores.

8 Las respuestas correctas se pueden a menudo encontrar en el texto cercano a la palabra que hay recuperar. De esta forma el candidato que adopta la correcta estrategia de resolver el test como si fuera un rompecabezas, puede tener ventaja sobre otro candidato que tenga una habilidad de la lengua similar.

9 El efecto rebote del C-test puede ser negativo si los profesores y alumnos emplean mucho tiempo en clase y en casa rellenando C-tests a costa de la práctica de otras destrezas más útiles y más parecidas a las actividades de la vida real. Kontra y Kormos (2006: 123)

McBeath (1989) afirma que el C-test carece de base teórica y Grotjahn (1987) comenta que las líneas maestras para diseñar el C-test son poco apropiadas para algunas lenguas, como el chino o el japonés, cuya escritura no es alfabética. Además, algunos investigadores también critican el hecho de que, debido al procedimiento de supresión, puede suceder que una palabra se suprima varias veces porque aparezca repetidamente en un texto, por ejemplo la palabra “and”, lo que haría que el test fuera muy fácil. Esto desautorizaría la afirmación de sus creadores que dicen que la supresión de las palabras no es decisiva.

4.5. Validación del C-test

4.5.1. Validez del constructo: ¿qué mide el C-test?

4.5.1.1. Identificación del constructo

Klein-Braley (1984, 1985 y 1994) intentó estudiar tanto las estrategias psicolingüísticas de los candidatos como si se podía predecir la dificultad de los C-tests basándose en las características de los textos. Ambos aspectos de su investigación son relevantes para la identificación del constructo. Según sus investigaciones, la predicción de los resultados venía definida principalmente por la media de palabras en cada frase y la ratio ”type-token”. Mientras que una ratio baja indica una escasa gama de vocabulario y una ratio alta indica lo contrario, la longitud de la frase suele ser un índice de la complejidad sintáctica de la misma y por lo tanto de la dificultad de procesamiento del C-test.

Una palabra puede ser difícil de recuperar debido a que el nivel de competencia de los candidatos es bajo y a que éstos no tienen en cuenta el contexto suficiente para excluir soluciones que pueden ser compatibles con el contexto local pero no con un contexto más amplio. Según ella, las características de los ítems que son fácilmente recuperables son:

- Alta frecuencia de la palabra en la lengua en general.
- La palabra forma parte de una frase hecha.
- La palabra tiene cierta similitud con otra palabra análoga en el primer idioma del candidato.
- Los pronombres posesivos.
- La palabra forma parte del campo semántico del tema del texto.

- La palabra se encuentra sin mutilar bien en el título del texto o bien antes o después del término que tienen que recuperar.

Para Sigott (2004), tanto la frecuencia de la palabra que tenían que recuperar como la clase de palabra que era tenían un efecto significativo en la dificultad de los ítems. Sin embargo, no se pudo determinar si la relación era casual o debido a la presencia de otras variables. Según él, nuestro conocimiento del constructo del C-test no ha variado mucho después de estas investigaciones, aunque podemos pensar que lo que mide el C-test es la habilidad de procesar vocabulario y sintaxis en textos que expresan ideas complejas.

4.5.1.2. El C-test como medida de la competencia general de la lengua

Se ha discutido ampliamente sobre la validez del C-test. Es decir sobre lo que realmente mide este test. Raatz (1984) realizó una investigación para contestar a la pregunta de lo que el C-test medía realmente. Estudió la correlación entre el C-test y las titulaciones de la escuela en alemán, en gramática, en inglés, y con otros tests de idiomas (validez convergente). Para ello extrajo información con diferentes tests que él elaboró, como tests de ortografía, etc., y de las notas que los profesores tenían. Llegó a la conclusión de que el C-test medía principalmente la competencia general de la lengua.

El estudio de Connelly (1997) apoyó las conclusiones de Raatz, ya que también afirmó que el C-test es un instrumento aceptable para “assess general language competence” y Klein-Braley afirma que los resultados obtenidos con los C-tests muestran relaciones significativas con otros aspectos del conocimiento de la lengua y de su uso.

> C-tests are authentic tests of the construct of general language proficiency. It has been possible to demonstrate that the results

> obtained with C-tests show meaningful relationships to other aspects of language knowledge and performance.
> (Klein-Braley, 1985: 101)

Esta afirmación concuerda con los resultados de Eckes y Grotjahn (2006) cuyos análisis también apoyan firmemente la conjetura de que el C-test es una medida de la competencia general de la lengua. A Dörnyei y Katona (1992) les interesaba también saber lo que medía el C-test y realizaron una investigación con alumnos de Hungría. Los resultados de su cuidadoso estudio estadístico mostró que:

1 El C-test parece correlacionar bien con otros tests de competencia de la lengua, lo cual prueba que es un instrumento de evaluación bueno y fiable. También es un método integrador de evaluación de la lengua que correlaciona altamente tanto con las destrezas orales como con las puntuaciones de vocabulario y gramática.

2 El C-test mide la competencia general de la lengua de forma más fiable que el *cloze* y es mucho más fácil de elaborar y corregir.

3 El C-test mide la habilidad general de la lengua a distintos niveles de competencia de forma precisa y eficiente

Sus resultados también indicaban que los C-tests que eran muy fáciles o muy difíciles para algún grupo en particular también tenían una validez aceptable.

4.5.1.3. El C-test como medida de la comprensión lectora

Alderson (2000a) acepta el C-test como una técnica válida para evaluar la comprensión lectora aunque es reticente a la hora de aceptar la Hipótesis de Competencia Unitaria:

> The Unitary Competence Hypothesis (UCH).... claiming that there was a unitary competence, or general language proficiency… is now generally discredited.
> (Alderson, 2002: 21)

El concepto de una competencia unitaria subyacente en las diferentes destrezas de una lengua extranjera se desarrolló en los años 1970. El principal argumento fue las correlaciones tan altas que existían entre los tests que median diferentes aspectos de competencia en una lengua extranjera. A pesar de la afirmación de Alderson de que esta hipótesis está ahora desacreditada, estas correlaciones tan altas se siguieron encontrando en otros estudios posteriores.

> High corelations have been found between set of scores from tests purporting to measure grammatical knowledge and sets of scores from tests purporting to measure lexical knowledge, and there has been little success in attempts to demonstrate that “grammar tests” and “vocabulary tests” tap fundamentally distinct aspects of linguistic knowledge. (Singleton y Singleton, 2002: 154)

El C-test se ha utilizado en otros idiomas distintos del inglés encontrándose correlaciones significativas entre el C-test y otros tests utilizados en esas lenguas. Esteban (2005), y Daller y Phelan (2006: 103) aseguran que el C-test mide la competencia general de un idioma y que la afirmación de Alderson tiene que ser modificada “in the sense that there is something resembling a common underlying profyciency at least in certain areas”. Esto puede ser cierto si tenemos en cuenta que el C-test no cubre todas las áreas de competencia de un idioma.

4.5.1.4. El C-test como medida del conocimiento del vocabulario

El conocimiento del vocabulario es un elemento importante en varias áreas del modelo de Bachman y Palmer (1996) lo mismo que para Read (2000)

quien describe la competencia léxica que está involucrada en el procesamiento de la tarea del *cloze* (incluyendo el C-test) como el conocimiento de palabras individuales y la habilidad de utilizar claves contextuales para determinar que palabras son apropiadas para un hueco.

> A cloze tends to make a very *embedded* assessment of vocabulary, to the extent that it is difficult to unearth the distinctive contribution that vocabulary makes to test performance.
> (Read, 2000: 115).

Singleton y Singleton (2002) defienden que con la ayuda del C-test se pueden investigar los procesos de adquisición léxica de un modo fiable y válido.

4.5.1.5. El C-test como medida del conocimiento de estructuras gramaticales

Para Eckes y Grotjahn (2006), el léxico y la gramática son importantes componentes de la competencia general de la lengua tal como los mide el C-test, pero advierte que lo que mide el C-test depende del nivel de competencia de los examinandos y de la dificultad del C-test.

> … their relative weight in contributing to performance on C-tests, and hence, the exact nature of what C-tests measure depends to some extent on the ability level of the examinees and on the difficulty of the C-test. (Eckes y Grotjahn, 2006: 316)

Dörnyei y Katona (1992) no consideran al C-test como un instrumento eficiente para evaluar la gramática. Por el contrario, Chapelle y Abraham (1990), en Babaii y Ansary 2001: 4) consideran que, debido a los textos cortos usados en el test y a la importancia de pistas en el entorno inmediato, es probable que los C-tests midan más la competencia gramatical que la textual.

4.5.1.6. El C-test como medida del grado de procesamiento del texto

Chapelle y Abraham (1990), y Babaii y Ansary muestran la importancia del conocimiento gramatical para hacer un examen de C-test pero indican que el C-test puede medir tanto las micro-características como las macro-características del lenguaje.

> C-test triggers both macro- and micro-aspects of the language, it conforms well to the principle of reduced redundancy which fundamentally emphasizes that both a global and a local knowledge are required to supply the missing elements in a distorted linguistic message. (Babaii y Ansary, 2001: 8)

Grotjahn y Stemmer (2002: 124) defienden que el hecho de que los estudiantes de una segunda lengua raramente utilicen el macro-contexto para hacer un C-test no significa necesariamente que el C-test no mida la competencia global de un idioma: ”L2 readers very often tend to perform mainly low-level processing”. También afirman que el C-test mide el componente receptivo de la competencia de la lengua académica. Sigott (2004, 2006) señala que la cantidad de procesamiento que realizan los alumnos a nivel textual probablemente dependa de la dificultad del pasaje en concreto sobre el que se ha elaborado el C-test o, lo que es lo mismo, del nivel de inglés de los alumnos. Él sugiere que los candidatos que pueden reconstruir un término haciendo uso tanto del contexto del pasaje completo (contexto textual) como del contexto reducido (contexto a nivel de frase) tienen mayor nivel de competencia en la lengua que los que solamente pueden reconstruir el término cuando tienen acceso al pasaje entero. Como resultado de esto, la proporción de palabras del C-test que se pueden recuperar valiéndose del contexto textual es probable que cambie en función del nivel de competencia de los candidatos. Consecuentemente, un mismo pasaje de C-test puede dar lugar a diferentes tests para alumnos de diferentes niveles de competencia de la lengua. Por lo tanto, se puede decir que el C-test tiene un “constructo fluido”, que cambia en

función de la habilidad de los candidatos y de la dificultad del pasaje. No se ha estudiado si este fenómeno ocurre también en otros modelos de tests de idiomas

Kontra y Kormos (2006) resumen parte de lo dicho anteriormente al asegurar que el C-test puede evaluar el conocimiento de elementos léxicos y estructuras gramaticales así como los componentes de la competencia comunicativa que definieron Bachman y Palmer (1996: 68) como: "knowledge of how individual utterances and sentences are organized" y también "knowledge of how sentences or utterances are organized to form a text". También afirman, basándose en los resultados de su estudio, que el C-test mide la competencia textual y que incluso han observado procesos de comprensión a nivel superior en el caso de estudiantes de nivel avanzado. Así pues el C-test mediría tanto los conocimientos léxicos y sintácticos como la competencia textual y global de una lengua para alumnos de niveles superiores y avanzados. Sin embargo, según sus estudios, el C-test parece proporcionar escasa información sobre el conocimiento que los alumnos tienen sobre morfología, ya que los candidatos elegían la palabra adecuada contando el número de letras que habían sido eliminadas de la palabra en lugar de tener en cuenta el contexto sintáctico.

Klein-Braley (1985) afirma que el C-test tiene una validez de constructo superior al *cloze*. Ella también nos demuestra que el C-test parece discriminar bien entre candidatos que se encuentran en diferentes etapas del aprendizaje de la lengua. Sin embargo, Jafarpur (1995) duda de su teórica superioridad sobre el *cloze* en cuanto a validez criterial se refiere (pp. 202-207).

Como hemos podido observar, no parece existir mucho acuerdo sobre los beneficios o los problemas del C-test o sobre la validez del constructo, es decir, sobre lo que el C-test puede o no puede medir. Sin embargo, la mayoría de los investigadores que han realizado estudios sobre el C-test están de acuerdo en que los conocimientos de gramática y léxico de los candidatos afectan a los resultados del test. También afirman que el C-test mide

principalmente la competencia general de una lengua y la de la comprensión lectora.

4.5.2. Validez aparente

La validez aparente del C-test no es muy buena. Jafarpur (1995) afirma que ni las opiniones de los profesores ni las de los alumnos son favorables a la apariencia o a la dificultad del C-test. Por otra parte, es interesante recalcar la discrepancia entre la dificultad percibida y la dificultad empírica de los tests. Bradshaw (1990) recogió información sobre las reacciones de alumnos españoles e italianos que realizaron un examen de clasificación que incluía tres pasajes de C-test. Los candidatos percibieron que el C-test era considerablemente más difícil que el test de gramática de elección múltiple o el test de comprensión lectora de respuestas abiertas. Sin embargo, se constató que no había diferencias entre las medias de las tres partes del examen.

Los C-tests se suelen considerar potencialmente frustrantes y existe considerable incertidumbre respecto a lo que realmente miden. Weir (1988: 53) cree que la validez aparente del C-test es baja, ya que los alumnos consideran irritante el tener que procesar textos que han sido excesivamente mutilados. Tendremos que esperar para comprobar si la realización del C-test por ordenador influirá en la validez aparente del test.

4.5.3. Validez de contenido

La validez de contenido no ha recibido mucha atención en la investigación sobre los C-tests. Puesto que los C-tests son considerados como tests de competencia general de la lengua, su contenido tiene que ser representativo de la lengua en su conjunto. Algunos investigadores comentan

la validez de contenido de los C-tests analizando hasta que punto la distribución de las clases de palabras afectadas por la mutilación es similar a la del texto original.

Raatz (1984) contestó a los que tenían dudas sobre lo que medía el C-test diciendo que puesto que las palabras mutiladas eran una muestra representativa de todas las palabras del texto, el C-test poseía validez de contenido:

> Since C-tests use several unsystematically selected texts, whose subject matter is as varied as possible, since in each case the mutilated words are a representative sample of all the words in the texts, and since, in addition, they are based on a linguistic theory, namely Oller´s construct of pragmatic expectancy grammar, we claim that C-tests are content valid tests. (Raatz, 1984: 125)

En un estudio crítico que comparaba los C-tests con los *clozes*, Jafarpur (1995) demostraba que variando el punto donde se empezaba a mutilar el texto y la frecuencia de las mutilaciones del mismo se obtenían tests en los cuales la distribución de palabras era completamente diferente a la del texto original. Sin embargo, se puede argumentar que la variación de esos dos parámetros viola las reglas de la construcción del C-test, por lo que los tests que resultan no se pueden considerar C-tests.

Apenas existe un test de idiomas que haya generado más controversia entre los expertos que el C-test cuando se trata de identificar el contenido de ítems individuales. Esto ha contribuido al escepticismo con el que muchas personas tratan a este test. Puesto que el C-test se construye sin tener que tener en cuenta en sus especificaciones ningún término lingüístico que distinga el contenido del test del método, la persona que elabora el test no tiene que realizar ninguna afirmación explícita acerca de lo que se supone que miden los diferentes ítems. Los intentos de validación del contenido, por lo tanto, nos conducen a intentar adivinar que destrezas o procesos utilizan los candidatos para recuperar una palabra mutilada.

Podemos resumir diciendo que, los estudios sobre la validez de contenido del C-test no parecen ser muy reveladores.

4.5.4. Validez concurrente

La mayoría de los estudios de validación han utilizado métodos de valoración externa, tales como correlacionar las puntuaciones del C-test con los resultados obtenidos usando otras medidas, por ejemplo Klein-Braley, (1997, 2002). Sin embargo, Alderson, Capham y Wall (1995) entre otros señalan que tales estudios necesitan ser tratados con cautela, ya que todos los tests estaban diseñados para medir diferentes aspectos de la lengua. A lo largo de los años, los C-tests han sido correlacionados con una gran variedad de distintos tests externos. Klein-Braley y Raatz (1984: 138) nos proporcionan una perspectiva general sobre los primeros estudios de validez concurrente. Los coeficientes de correlación estaban entre 0,40 y 0,95, indicando un considerable grado de solapamiento en los constructos que se medían.

Sigott (2004) también nos presenta un resumen de la información que se ha publicado más recientemente sobre la validez concurrente de los C-tests. Mientras que la variedad de criterios con los que los C-tests han producido correlaciones significativas es impresionante, los coeficientes de correlación de los C-tests con tests que se supone que miden constructos similares varía considerablemente. Por ejemplo, los coeficientes de correlación entre el C-test y tests que miden la comprensión lectora de un candidato varían entre 0, 29 y 0,86. Lo mismo ocurre con las destrezas de comprensión oral, expresión oral y estructuras gramaticales. Sin embargo, los coeficientes de correlación entre el C-test y los tests que miden la competencia global de la lengua ofrecen valores de medianos a altos. Estos resultados son consistentes con los estudios analíticos en los cuales se observa que el C-test contribuye en gran medida a un factor que se ha llamado “competencia general de la lengua”.

4.6. Factores a tener en cuenta al elaborar un C-test

En general, se utilizan textos fáciles en la construcción del C-test. Sin embargo, Babaii y Ansary (2001) sugieren que los textos deben plantear un reto a los examinandos para que se active el procesamiento del texto a nivel macro. Los que han desarrollado el método del C-test sugieren que se deben ordenar los textos en orden ascendente de dificultad para, de acuerdo con Spolsky (1981), mejorar la distribución de las puntuaciones que se obtienen antes de ser procesadas.

Puesto que no es muy fácil decir de antemano si los candidatos encontrarán un texto más difícil que el otro, Klein-Braley (1984) intentó estudiar si es posible predecir la dificultad tanto de los textos como de los C-tests. Ella sugiere que el valor de la media del C-test debe estar alrededor del 50% para asegurar la máxima diferenciación entre los sujetos y aconseja ordenar el test en orden ascendente de dificultad, empezando con un texto fácil, que ella denomina "icebreaker" y que ayuda a los candidatos a entender en que consiste el método del C-test, y terminando con uno realmente difícil.

> It is important for the icebreaker to be so simple that every test subject understands exactly what the C-principle demands from him or her. (Klein-Braley, 1984: 98)

Para su estudio, empleó el mismo enfoque que el de los investigadores de la legibilidad, es decir, intentó relacionar las características estadísticas de los textos – gama de vocabulario, longitud de las frases, frecuencia de las palabras de léxico, la proporción type-token, etc., y las dificultades del C-test determinadas empíricamente.

No alcanzó ningún resultado decisivo aunque cree que la dificultad de procesar el C-test decrece de forma lineal a la vez que las personas del primer idioma se hacen mayores. Esto es debido a que las estrategias para procesar

la lengua son más eficientes a medida que la edad de los individuos aumenta. Esto es lógico si tenemos en cuenta que para procesar un texto se usan tanto los conocimientos de la lengua como los del mundo real, y que ambos conocimientos suelen aumentar con la edad.

Elementos con índices de baja dificultad tienen un efecto tope en los C-tests. Para evitar eso, algunos investigadores, tales como, Sigott y Kobrel (1996), o Cleary (1988) sugieren aumentar el grado de dificultad de los textos bien dejando solamente la primera letra sin suprimir en las palabras en las que se suprimen letras, o bien suprimiendo la primera mitad de la palabra en lugar de la última (left hand deletion).

Cleary suprimió las letras de la izquierda de las palabras que no tenían terminaciones morfológicas en un texto de un C-test de inglés y comparó las medias con la versión de C-test estándar. Los resultados fueron que las medias de la versión experimental fueron considerablemente más bajas que las del C-test estándar. Por lo tanto, la supresión de letras por la izquierda aumenta el grado de dificultad del C-test.

4.7. Usos del C-test

La mayoría de los investigadores parecen estar de acuerdo, según se ha visto en este capítulo, en que los C-tests son tests de competencia o de *proficiency* y se califican con una sola puntuación, calculada al sumar las puntuaciones obtenidas en los subtests. Esta puntuación representa la posición global del individuo en el continuo de competencia de la lengua. Los C-tests se pueden utilizar para todas las personas que estén aprendiendo un idioma, incluidos los niños nativos.

Como ya se dijo anteriormente, la puntuación puede interpretarse con relación a las puntuaciones del resto del grupo o en términos absolutos, como

una nota de corte para un nivel determinado, o como un indicador de progreso hacia la competencia de un hablante nativo. De acuerdo con esto, los C-tests, podrían utilizarse como pruebas de nivel, pruebas de selección, o pruebas de progreso. Sin embargo, los C-tests, puesto que no pueden ser manipulados para seleccionar estructuras específicas o un grado de dificultad determinado, no se pueden usar para pruebas de diagnóstico.

Uno de los exámenes a gran escala que tenemos que realizar en la escuela todos los años es el examen de clasificación de las personas que quieren asistir a nuestra institución como alumnos oficiales. Para ello tenemos que determinar la competencia de la lengua que los solicitantes tienen para así poder distribuirlos, de acuerdo con su nivel, en los seis cursos en los que se divide la enseñanza de inglés como lengua extranjera en la EOI.

Norris (2006) demostró que los C-tests pueden ser utilizados como exámenes de clasificación de los estudiantes que llegan nuevos a una institución y hay que decidir a que curso deben incorporarse. Su investigación buscaba extender y explorar el desarrollo y uso del C-test para clasificar a los alumnos y colocarlos en el nivel que les correspondía de acuerdo con el currículo de un colegio alemán en Estados Unidos. Según sus investigaciones, el C-test distingue entre alumnos de diferente nivel curricular y puede agruparlos por nivel de competencia.

Daller y Phelan (2006) afirman que el C-test mide la competencia general de una lengua y que es una técnica válida tanto para ser utilizada como examen de clasificación como para medir el progreso de los estudiantes a lo largo del curso. También comprobaron que existía una correlación significativa entre el C-test y los subtests que forman parte del examen “Test of English for International Communication” (TOEIC®) lo que demuestra que el C-test tiene una alta validez para predecir el comportamiento de los alumnos en ese curso de lengua extranjera.

4.8. El vocabulario y el C-test

Meara y Fitzpatrick (2000: 21), Read (2000) y Nation (2001) consideran que los ítems de tests similares al C-test trazan una fina línea entre destrezas productivas y receptivas, ya que la recuperación de una palabra depende de la comprensión receptiva de las palabras del contexto que la rodean. Las palabras no mutiladas proporcionan las claves necesarias para restaurar la palabra mutilada. Además de inferir el significado de la palabra que tienen que restaurar, los candidatos tienen que conocer la ortografía, las partes de la palabra (conocer las reglas de la formación de palabras), la estructura de la frase, la organización del párrafo, etc. Es decir, que el candidato tiene que usar varias clases de conocimiento de la lengua para recuperar correctamente la palabra léxica mutilada. Por lo tanto, el terminar con éxito una tarea de c*loze*, incluido el C-test, implica construir el significado de las palabras en contexto usando el conocimiento sintáctico y esquemático, tanto para las palabras mutiladas como para las palabras del texto que las rodean Nagy (1997). Los lectores utilizan toda esta información contextual para decidir que palabras están semánticamente relacionadas con el contexto y que palabras no lo están Corrigan (2007).

Lee (2008) y Nation (2001) afirman que para recuperar palabras de léxico en un C-test se requiere el conocimiento de las categorías sintácticas de las palabras y sus propiedades sintagmáticas, es decir, se requiere conocer el vocabulario con profundidad. También se requiere conocer cuándo ciertas palabras aparecen juntas con mucha frecuencia formando colocaciones típicas, expresiones idiomáticas o frases léxicas. Los procesos que se activan para producir palabras son similares a los que se usan en la lectura, aunque son más conscientes en la producción del lenguaje, ya que siempre *el recordar* es más difícil que *el reconocer*, bien sea la forma o el significado de una palabra.

Singleton y Little (1991) usaron el C-test como instrumento de investigación para proporcionar evidencia sobre la naturaleza del conocimiento

del vocabulario en una segunda lengua. Para ello, administraron un C-test a alumnos universitarios y estudiaron cuidadosamente las respuestas incorrectas, llegando a la conclusión de que los estudiantes eran capaces de utilizar las claves semánticas del texto y adivinar cuál debía ser el significado que se necesitaba expresar con la palabra mutilada. De lo que carecían los estudiantes era del conocimiento de la forma exacta de la palabra. Esto significaba que los alumnos usaban bien una forma posible pero incorrecta o bien una forma que no existe. Basándose en estos resultados Singleton y Little afirmaron que el C-test era un instrumento útil para medir el conocimiento del vocabulario de los estudiantes de una segunda lengua y que las palabras que estos estudiantes aprendían se organizaban basándose en la semántica y no en la fonética.

Chapelle (1994) retó la segunda afirmación de Singleton y Little (1991) argumentando que no se sabía si el proceso que los estudiantes utilizaban para recuperar cada palabra era consciente o automático. Ella sugirió que si los profesores e investigadores querían utilizar el C-test para evaluar el vocabulario, se debería abandonar la ratio fija y elegir solamente palabras léxicas que podrían ser restauradas utilizando únicamente claves contextuales.

Nation (2001) considera que el recuperar la palabra que falta es una subdestreza de la comprensión lectora y depende principalmente del nivel de competencia que tengan nuestros alumnos. Recomienda que se debe trabajar con textos en los que al menos el 95% de las palabras sean conocidas por los estudiantes para que puedan acceder a las pistas que el texto les proporciona. Como ya vimos en el capítulo 3, no todas las palabras se adivinan con igual facilidad, ya que no todas ellas ofrecen las mismas claves. Según Nation (2001) los adjetivos no se relacionan generalmente con otras palabras y son más difíciles de restaurar que los nombres y los verbos

4.9. Conclusiones

Desde el punto de vista de la construcción, los C-tests ofrecen ciertas ventajas. Los textos, excepto raras excepciones, se pueden obtener de los materiales de lectura que sean apropiados para el grupo que se desea examinar. Una vez que tenemos los pasajes, es muy fácil convertirlos en C-tests. Solamente hay que asegurarse de que palabras tales como los nombres propios, los números y las abreviaciones no resulten mutiladas.

Por una parte, Los C-tests no han disfrutado de muy buena prensa en cuanto a validez aparente se refiere. Se han descrito a menudo como frustrantes para los alumnos que se preguntaban que estaría midiendo realmente este test. El C-test presenta también problemas con respecto a su contenido. Esto es debido a la generalidad de la teoría en la que se basa y a que no se requiere que el constructor del test tenga que explicar explícitamente que es lo que mide un ítem determinado.

Por otra parte, los C-tests tienen, en general, coeficientes de consistencia interna elevados incluso con tests que no han sido pilotados con anterioridad. También ofrecen generalmente resultados muy aceptables en términos de validez concurrente, al obtenerse correlaciones bastante altas con una amplia variedad de tests que miden diferentes habilidades de la competencia global de la lengua. Existen evidencias distintas en cuanto a la validez del constructo se refiere. Los C-tests pueden medir el aumento de competencia lingüística y desde el punto de vista psicométrico, considerando los pasajes que forman el C-test como “super-ítems”, el C-test tiende a aproximarse a la unidimensionalidad. Esto está de acuerdo con los altos valores de consistencia interna que se obtienen.

Además, los factores que pueden determinar la dificultad de los textos, por ejemplo la longitud de las frases o la ratio “type-token”, son tan vagos y generales que contribuyen poco a nuestro entendimiento del constructo del

C-test. Además, existen multitud de factores que pueden determinar que un individuo recupere o no la palabra mutilada en un texto por lo que los intentos de predecir la dificultad de ítems individuales con vistas a comprender la naturaleza del constructo no han sido fructíferos.

Quizás el tema más importante, en cuanto a la validez del constructo se refiere, ha surgido de estudios centrados en analizar los procesos mentales involucrados en resolver un C-test y en saber si para ello los candidatos tienen que hacer uso de distintos niveles de procesamiento del texto, es decir, si tienen que utilizar contextos que van más allá del entorno inmediato de los huecos. Las investigaciones que se han realizado sobre este aspecto son contradictorias. El estudio de Sigott (2004) proporcionó alguna clarificación al afirmar que el nivel de procesamiento de los textos que exige el C-test depende del nivel de competencia de la persona que intenta resolverlo, por lo que puede darse a todos los niveles del discurso desde el nivel de la palabra hasta el nivel del texto. Desde este punto de vista, el mismo C-test podría originar diferentes tests para diferentes niveles de competencia, lo cual podría afectar a la interpretabilidad de los los resultados del C-test.

Mientras se arroja más luz sobre este tema, el C-test usado en exámenes a gran escala debe someterse al mismo proceso de validación que cualquier otro modelo de examen.

Capítulo 5

LA EVALUACIÓN EN LA ESCUELA OFICIAL DE IDIOMAS

5.1. Introducción

La Escuela Oficial de Idiomas (EOI) es una institución pública de ámbito estatal que el próximo año celebrará su primer centenario, ya que fue fundada en el año 1911 en Madrid y después se extendió por toda España. En los diversos centros de la EOI se imparten diferentes idiomas cuya su enseñanza no es obligatoria y va dirigida principalmente a la educación de la población adulta. La Escuela más grande se encuentra en Madrid y en ella se enseñan 22 idiomas, incluidos todos los de la Unión Europea y también, desde 1985, todas las lenguas del Estado español.

Este proyecto ha originado muchas veces la admiración y la envidia de los países que pertenecen a la Unión Europea, ya que es un proyecto único en Europa y refleja exactamente la filosofía del Consejo de Europa que cree que la educación de una persona continúa a lo largo de toda su vida, y ha impulsado y favorecido de forma especial el que los europeos, que son conscientes de que comparten muchas tradiciones comunes pero que también poseen diversidad de culturas y de lenguas, lleguen a hablar al menos tres idiomas distintos. Ello favorecerá la movilidad de estudiantes y profesionales dentro de la Unión Europea.

Se utiliza la metáfora de que Europa es nuestra casa común y que las lenguas son la llave que abre el mundo cultural de Europa. Los idiomas juegan un papel clave en la apertura de mundos culturales.

Eckes et al. afirman que el aprender nuevos idiomas proporciona acceso a otras comunidades lingüísticas, a sus valores y a diferentes modos de entender el mundo. Todo esto contribuye a superar las distancias sociales, prejuicios y miedos que puedan existir.

> Learning foreign languages provides access to other language communities, their values, norms and perspectives of the world,

> and thus contributes to overcoming social distance, prejudice and fears, which may exist prior to, or may arise during, the process of integration. (Eckes et al. 2005: 356).

La publicación del Marco Común Europeo de Referencia para los idiomas (MCER) en Europa y la cada vez más extendida adopción de su compañero el Porfolio Europeo para los Idiomas (ELP) ha reanudado el debate sobre la evaluación, y nos ha concienciado sobre la necesidad de obtener uniformidad en la certificación de las lenguas modernas y adaptar nuestros exámenes al MCER, especialmente usando sus *escalas* como un instrumento. La política de que las certificaciones que se expidan en la Unión Europea sean homogéneas ha originado un mayor entendimiento público, el perfeccionamiento y el compartimiento de los métodos de evaluación y corrección por todos los países que forman la Unión Europea.

La EOI no ha sido ajena a estos cambios que se han reflejado en la adopción de nuevas técnicas para evaluar y nuevos hábitos para corregir los tests dando más importancia a la validez, la fiabilidad y la credibilidad de nuestros exámenes. Se han establecido nuevos procedimientos para elaborar y corregir los exámenes y se han realizado numerosas sesiones de estandarización y puesta en común de los criterios que se iban a tener en cuenta para evaluar a nuestros alumnos, teniendo siempre en cuenta el MCER.

5.2. Historia reciente de la evaluación en la EOI

Como es natural, la evaluación del inglés en la Escuela Oficial de Idiomas ha variado considerablemente a lo largo de estos cien años. Yo voy a limitarme a describir la evolución que ha sufrido durante los últimos 25 años, ya que ello forma parte de mi experiencia profesional en esta institución. Debo subrayar que la preocupación por la evaluación de la competencia del inglés, tanto de los alumnos oficiales como de los libres, ha ido en aumento. Ya no nos sirve cualquier examen que se prepare sin ninguna base teórica. Queremos

que los tests que se utilicen en la EOI sean fiables, válidos y comunicativos, es decir, que reflejen actividades de la vida real.

La enseñanza de idiomas que se impartía en la EOI se distribuía en 5 cursos y se realizaban exámenes varias veces al año: Un examen de clasificación en septiembre para las personas que querían estudiar en la Escuela como alumnos oficiales, tres exámenes refenciados a un criterio (en febrero, junio y septiembre) para los alumnos de primero, segundo y cuarto curso y exámenes de competencia tres veces al año también en febrero, junio y septiembre para los cursos tercero y quinto. A estos últimos alumnos se les daba un certificado oficial si aprobaban los exámenes en junio o septiembre.

La evaluación en la EOI ha cambiado a lo largo del tiempo. Algunos años atrás, se creía que examinando a los alumnos de gramática de la forma más exhaustiva posible se obtenía la mejor base para predecir la habilidad de la lengua. Se usaban técnicas tales como, el dictado, la traducción directa e inversa, el parafraseado y las preguntas de elección múltiple para evaluar el vocabulario, la gramática y la fonética.

Los candidatos, excepto los alumnos externos, eran examinados por sus propios profesores y la doble corrección para pruebas subjetivas no existía.

Los profesores, que preparaban sus propios exámenes, tenían poca o ninguna experiencia en la elaboración de los mismos. Además, los tests no eran comprobados nunca por ninguna persona especializada en la elaboración de pruebas. Como estos exámenes ofrecían poca confianza a los profesores, el departamento decidió cambiarlos. Se acordó cambiar las técnicas de examen y también que los exámenes fueran unificados, es decir, todos los cursos de un mismo nivel tendrían el mismo examen final, independientemente de los profesores que lo hubieran impartido.

Se empezó analizando los tests de competencia global (voy a concentrarme en estos, ya que los demás simplemente les siguieron). El departamento de inglés consideró conveniente que se examinara a todos los

alumnos, oficiales y libres, de las cuatro destrezas, es decir, de la comprensión escrita, la comprensión oral, la expresión escrita y la expresión oral, esperando que el efecto rebote que se produciría en la enseñanza de la lengua fuera positivo. Se cree que los tipos de tests que se utilizan en los exámenes animan a los profesores y a los alumnos a su práctica, tanto en clase como fuera de ella. Con ello se quería conseguir un aumento de la práctica de las destrezas orales en clase. Bachman y Palmer (1996: 75), sin embargo, no consideran que la lengua conste de cuatro destrezas. Ellos creen que el uso de la lengua se origina "in the performance of specific situated language use tasks". Es decir, que cuando hacemos algo lo hacemos "for a particular purpose, in a particular setting". Además, muchas tareas implican más de una destreza. Por ejemplo, lo que ellos llaman "correspondence" incluye comprensión lectora y expresión escrita (p.128).

Para minimizar la influencia del método de examen en la actuación de los alumnos, se decidió usar dos técnicas para evaluar cada destreza. Los tests deberían ser lo que Bachman (1990: 77) llama 'performance tests', en los cuales "the test taker´ test performance is expected to replicate their language performance in non-test situations". Esto debería ser así especialmente cuando se evaluara la expresión oral y la expresión escrita. Así se podría relacionar más fácilmente la actuación del candidato en el examen con el uso que el mismo haría de la lengua en una situación de la vida real, con lo cual aumentaría la validez aparente de los exámenes.

Los exámenes unificados los elaboraban, en un principio, los coordinadores del Departamento de Inglés, que en general no tenían ni experiencia ni formación en la preparación de exámenes a gran escala. La corrección seguía corriendo a cargo de los profesores que habían impartido cada curso, lo mismo que la nota final de los exámenes. Como los tests los corregía un solo profesor, las pruebas subjetivas como son la expresión escrita y la expresión oral no recibían doble puntuación.

Los tests de comprensión oral consistían en dos pasajes diferentes que eran leídos por el profesor que vigilaba cada aula. Ello originaba que las

lecturas no fueran homogéneas, ya que cada profesor podía leer a distinta velocidad o con distinto acento. Las dos tareas consistían en contestar a un test de elección múltiple y a varias preguntas de respuestas abiertas. Más tarde, para solventar estas diferencias, dos profesores nativos grababan los pasajes, de forma que todos los candidatos los oían exactamente en las mismas condiciones, es decir, a la misma velocidad, con el mismo acento y por la misma persona.

El examen de comprensión lectora consistía en un test de preguntas abiertas y en otro de elección múltiple, que medía principalmente los conocimientos de gramática, vocabulario y fonética. Sin embargo, el profesorado mostraba su descontento con el test de preguntas abiertas, ya que aunque los profesores tenían instrucciones de no penalizar las respuestas que mostraban que el alumno había comprendido el texto, aunque tuviera errores gramaticales o de ortografía, la realidad nos enseñaba que no todos los profesores seguían esas instrucciones por lo que continuaban penalizando errores que no eran de comprensión lectora. Se demostró que la corrección de algunos profesores podía ser sesgada, inaceptablemente variable y poco fiable. Es decir que las respuestas por escrito de los alumnos podían contaminar la medida de la comprensión lectora de los mismos, y necesitábamos tener una idea clara de la habilidad de lectura de los alumnos de la forma lo menos contaminada posible.

Se sugirieron y se probaron varias alternativas, tales como, el resumen, diversos tipos de *cloze* o las preguntas de elección múltiple. Como ya se ha visto en el capítulo 3 dedicado a la evaluación de la comprensión lectora, todas estas técnicas tienen problemas, algunos de los cuales las convierten en poco fiables.

La expresión escrita constaba de dos redacciones que, como ya se ha dicho, eran corregidas por el profesor que había impartido el curso. Esto quiere decir que la corrección no ofrecía toda la fiabilidad que debiera.

La expresión oral consistía en una interacción entre el profesor y sus alumnos, con lo que también carecía de la fiabilidad necesaria.

Mientras tanto, en el Segundo Congreso Nacional de Escuelas Oficiales de Idiomas, en Noviembre del 2001, se vio la necesidad de usar procedimientos estandarizados por autonomías para aumentar la validez, fiabilidad y credibilidad de los exámenes que estábamos utilizando y relacionarlos con el del Marco Común Europeo de Referencia para las Lenguas (MCER). Se decidió utilizar solamente un examen que consistía en una batería de tests para cada Certificado que se expedía en las Escuelas Oficiales de Idiomas. En aquella época se hacían exámenes de dos certificados, el llamado de Ciclo Elemental, que certificaba un nivel intermedio semejante al nivel de *First Certificate* de Cambridge, y el de Ciclo Superior, que certificaba un nivel avanzado semejante al nivel de *Proficiency* de la Universidad de Cambridge.

Como ya llevábamos varios años de duro trabajo intentando unificar nuestros exámenes durante los cuales se habían probado varias técnicas de examen, aceptando y rechazando algunas de ellas, fue más fácil llegar a un acuerdo sobre las tareas que debían formar parte del nuevo examen que iba a ser compartido por todas las Escuelas de la Comunidad de Madrid. El nuevo examen basado en materiales orales o escritos que procedían de fuentes reales, no de libros de texto, y en tareas semejantes a las que se realizan en la vida real, era un hecho. La Comunidad de Madrid nos dio apoyo y algunos profesores fueron liberados para dedicarse a la elaboración de dichos exámenes de certificado.

Las Escuelas de Idiomas de la Comunidad de Madrid compartían los mismos currículos, los mismos programas, los mismos exámenes y las mismas especificaciones con descripciones detalladas de la administración y corrección de las pruebas. Se prestó especial atención a la corrección de los tests subjetivos, sobre todo los tests de expresión escrita y oral. Para ello se realizaron numerosas sesiones de estandarización entre las personas que elaboraban los exámenes, los profesores que tenían que administrarlos y corregirlos y personas invitadas, expertas en evaluación, que procedían de

distintas universidades europeas, como por ejemplo, de la Universidad de Lancaster, UK. Por supuesto, los tests subjetivos se corregían por al menos dos personas distintas.

En el curso 2007-2008 se cambiaron los planes de estudio de las Escuelas Oficiales de Idiomas para adaptar nuestros niveles a los del Marco Común Europeo de Referencia para las Lenguas (MCER). Ahora existen seis cursos, en lugar de los cinco anteriores y tres certificados en lugar de dos. El certificado de nivel básico equivale al nivel A2, el certificado de nivel intermedio al B1 y el certificado de nivel avanzado al B2 del MCER respectivamente.

Aunque los exámenes y nuestros alumnos alcanzan niveles superiores al B2, sin embargo, no nos es posible certificar niveles superiores al B2. En este momento se está luchando para lograr que las Escuelas de Idiomas puedan certificar el nivel C, sobre todo de las lenguas estatales e idiomas afines. En algunas autonomías, como Cataluña y el País Vasco, ya pueden impartir y certificar el nivel C.

Los métodos de evaluación que forman parte del actual certificado de Nivel Intermedio no han variado sustancialmente de los que componían el anterior certificado de Ciclo Elemental. Se siguen evaluando las cuatro destrezas y solamente se ha añadido un nuevo test en la comprensión de la lectura, en el cual, se les pide a los candidatos que localicen información específica en textos relativamente extensos procedentes de diferentes fuentes, y con un tiempo limitado.

La fiabilidad en la corrección sigue siendo una de nuestras principales preocupaciones. Se continúa todos los años con las sesiones de estandarización, basadas en exámenes reales, para corregir posibles desviaciones y se continúa también invitando a personas que creemos que nos pueden ayudar, debido a su experiencia en este campo, a mejorar la calidad de nuestros exámenes. En septiembre del año 2009, por ejemplo, Brian North, perteneciente a Eurocentres Foundation e investigador clave involucrado en la creación del MCER, tuvo la amabilidad de aceptar nuestra invitación para

trabajar con nosotros y asesorarnos en el uso de las escalas que describen los distintos niveles del MCER. Así mismo, sus sugerencias y consejos nos iluminaron sobre el grado de competencia que deberíamos esperar de los alumnos en los distintos niveles que ahora se están impartiendo en la EOI.

Las Escuelas Oficiales de Idiomas, como una de las instituciones que administran certificados oficiales en España, está obligada a seguir las directrices del MCER (ver Figueras et al., 2005) para asegurar que los niveles de competencia tienen el mismo significado en cualquier lugar de la Unión Europea. Aunque de acuerdo con Little (2005) el MCER presenta varias limitaciones y puede ser mejorado, sin embargo, pensamos que la dirección es la correcta y que el futuro para la EOI puede ser muy prometedor si recibe el apoyo y la cooperación institucional que se necesita para realizar todos los cambios necesarios y formar a todas las personas involucradas en la reforma de los sistemas y prácticas de elaboración de exámenes y evaluación de nuestros alumnos.

5.3. Técnicas de evaluación usadas en el certificado de Ciclo Elemental.

A pesar de que ahora tenemos otro plan de estudios, voy a describir los tests que se usaban en la EOI en el año 2007 para evaluar a los alumnos de tercer curso que se presentaban al certificado de nivel intermedio llamado Ciclo Elemental. Ello es debido a que esta investigación está basada en los tests que se utilizaron ese año en la EOI Jesús Maestro de Madrid conjuntamente con el C-test que se elaboró para estudiar la conveniencia de sustituir el *cloze* test, que formaba y todavía forma parte de la batería de tests que componen el examen de comprensión lectora, por la técnica del C-test.

Las características generales de las diferentes técnicas usadas en el ciclo elemental para evaluar las cuatro destrezas de la lengua eran las siguientes:

5.3.1. Expresión escrita

La evaluación se realiza a partir de dos tareas auténticas o verosímiles, de interacción o de expresión, claramente contextualizadas y de diferente longitud. El candidato debe demostrar que es capaz de escribir textos sencillos sobre temas generales de diversa tipología (cartas y mensajes personales, cartas formales tipificadas, instrucciones, solicitudes, cuestionarios, breves informes, descripciones o relatos) en los que se solicita o transmite información, se describen o narran acontecimientos, hechos imaginarios, sueños, deseos, reacciones y sentimientos, se justifican brevemente las opiniones y se explican planes o proyectos.

Las tareas propuestas deben medir la interacción y la expresión escrita del candidato. Cada una de las dos tareas debe tener un objetivo diferente de producción escrita y género textual distinto. Los tests de expresión escrita se corrigen por dos profesores que usan unas escalas analíticas que adjudican notas de forma separada a los diferentes componentes del constructo que se desea medir. Entre los componentes que se evalúan se encuentran:

a. Eficacia comunicativa: comprensibilidad, cumplimiento de las funciones esperadas y adecuación sociolingüística.

b. Capacidad discursiva: coherencia de las ideas, organización y desarrollo.

c. Uso de la lengua: exponentes lingüísticos, recursos formales de cohesión y flexibilidad.

d. Corrección formal: gramática, cohesión discursiva, vocabulario y ortografía.

La calificación de cada tarea se obtiene mediante la suma aritmética de las puntuaciones obtenidas en cada uno de los criterios. Como ya se ha dicho, se han realizado sesiones de estandarización para asegurar la correcta aplicación de las escalas por todas las personas que corrigen y evitar así que exista disparidad de criterios.

5.3.2. Expresión oral

La evaluación se realiza a partir de dos tareas en las que se proponen situaciones auténticas o verosímiles de interacción con otras personas y de exposición, con o sin apoyo visual.

En la primera tarea se trata de interactuar con otro candidato partiendo de una situación común con pautas concretas referidas a la situación y a los objetivos comunicativos, pero sin imponer identidades ficticias a los candidatos. El candidato debe demostrar, en interacción y como hablante, que es capaz de interactuar y de expresarse en situaciones incluso menos habituales y sobre temas concretos o abstractos para relacionarse, intercambiar opiniones e información detallada.

La segunda tarea se trata de un monólogo sostenido partiendo de unas pautas concretas. Estas pautas son orientaciones referidas a la situación y a los objetivos comunicativos, pero no imponen identidades ficticias ni opiniones específicas que los candidatos tengan que defender forzosamente. El candidato debe narrar y describir experiencias, sentimientos y acontecimientos, presentar un tema conocido y justificar brevemente las propias opiniones en un registro estándar de formalidad e informalidad.

La expresión oral de los candidatos se evalúa, lo mismo que la escrita, por dos profesores con arreglo a los siguientes criterios:

a. Eficacia comunicativa: comprensibilidad, cumplimiento de las funciones esperadas, precisión, adecuación sociolingüística.

b. Capacidad interactiva y discursiva: reacción y cooperación, coherencia de las ideas, organización, desarrollo relevante y suficiente.

c. Uso de la lengua: recursos lingüísticos, elementos formales de cohesión y fluidez.

d. Corrección formal: gramática, cohesión discursiva, vocabulario y pronunciación.

La calificación de la expresión oral se obtiene mediante la suma aritmética de las puntuaciones obtenidas en cada uno de los criterios anteriormente mencionados y que vienen reflejados en las tablas que los profesores utilizan para la evaluación de cada candidato.

Al tratarse de una prueba subjetiva, se realizan también sesiones de estandarización en las que se analizan grabaciones de actuaciones reales de alumnos en exámenes anteriores. Estas sesiones nos ayudan a unificar criterios y a utilizar las tablas que se manejan en la evaluación de forma consistente por todos los profesores, aumentando así la fiabilidad de las notas obtenidas por los candidatos.

5.3.3. Comprensión oral

La evaluación se realza a partir de textos orales auténticos o verosímiles, explotados por primera vez para la ocasión y procedentes de

fuentes tales como la radio, la televisión, grabaciones no comerciales, etc. Los textos orales pueden incluir noticias, previsiones del tiempo, mensajes telefónicos, anuncios públicos y publicitarios, conversaciones de carácter informal y reportajes o entrevistas sobre temas generales.

El candidato debe demostrar que es capaz de identificar las intenciones comunicativas, el tema, las ideas principales, los detalles más relevantes, seleccionar la información pertinente y captar el registro de los textos claramente estructurados (informaciones, instrucciones y explicaciones sencillas, indicaciones detalladas, noticias, mensajes telefónicos, documentales o programas en los que se narra o se presenta un tema, debates y entrevistas) sobre temas generales o de su especialidad, sobre los que pueda formular hipótesis de contenido, emitidos de forma relativamente lenta y clara, en registros formales o informales estándar y con posibilidad de volver a escuchar o aclarar dudas.

Las tareas de comprensión oral miden la comprensión global, la de ideas principales, la de ideas secundarias más destacadas y la de detalles. Consta de tres actividades concretas distintas, cada una con un objetivo diferente de comprensión. Las tareas concretas que realizaron los alumnos que forman parte de esta investigación fueron tres:

- Relacionar 5 textos cortos con sus títulos o con frases que aludan al sentido global. Se proporcionaron 8 títulos que incluían los 5 títulos correctos más tres distractores. El objetivo es relacionar las intenciones comunicativas o el tema de los textos cortos.

- Test de elección múltiple, con tres opciones de respuesta para cada ítem. El objetivo es comprender las ideas principales y secundarias más destacadas.

- Rellenar un formulario o cuaderno de notas con datos o información específica. El objetivo es el de seleccionar información específica.

El examen de comprensión auditiva lo corrige un solo profesor de acuerdo con la clave que se le proporciona.

5.3.4. Comprensión de lectura

La evaluación se realiza a partir de textos escritos auténticos o verosímiles, explotados por primera vez para la ocasión, adaptados como y cuando se considere pertinente, de tipología diversa y procedentes de fuentes tales como prensa, internet, publicaciones de instituciones oficiales o entidades públicas o privadas, comerciales, etc.

El candidato tiene que demostrar que es capaz de identificar las intenciones comunicativas, el tema, las ideas principales, los detalles más relevantes, el hilo argumental y las conclusiones de textos claros y bien organizados sobre temas generales o relacionados con su especialidad (mensajes y textos de relación social, anuncios de trabajo o publicitarios, folletos turísticos o comerciales, instrucciones, noticias, relatos y artículos de opinión o de información no especializados), así como localizar información procedente de distintas fuentes en los mismos tipos de textos.

Las tareas concretas que se les pidió a los alumnos para este trabajo de investigación fueron tres:

- Headings: Emparejar textos cortos con epígrafes que hagan referencia a su contenido. Se proporcionan 7 textos cortos y 10 epígrafes (los 7 epígrafes correctos más 3 distractores). El objetivo es identificar las intenciones comunicativas o el tema de los textos cortos, es decir, se trata de evaluar la comprensión global del texto.

- *Cloze*: consistió en rellenar huecos con un banco de ítems. El *cloze* se basaba en un solo texto con 20 huecos y el banco de ítems constaba de 25 palabras, ya que a los ítems que había que utilizar para rellenar el texto se añadían 5 distractores. El objetivo es comprender información sobre detalles relevantes, es decir, que se trata de evaluar la comprensión detallada del texto.

- Opción múltiple: con este test se trata de evaluar la comprensión de la información principal o secundaria más destacada.

En todos y cada uno de los tests se proporcionó la respuesta a un ítem que servía como ejemplo. Como se ha demostrado que un test de gramática añade escasa información extra a la batería de tests de competencia lingüística, no se vio la necesidad de evaluar la gramática de forma separada de la comprensión de lectura. A cada destreza se le adjudica una sola nota que es la suma de las puntuaciones de las diferentes tareas de que consta el test. Todas las destrezas tienen el mismo peso, y la puntuación mínima para aprobar es el 60% de la nota total de la destreza.

Las tareas tanto de comprensión lectora como de comprensión auditiva son prácticamente 100% objetivas, en cuanto a la corrección se refiere. Por lo tanto la fiabilidad entre los profesores que corrigen es muy alta. Se intenta utilizar textos basados en temas conocidos y frecuentes, ya que el conocimiento del tema facilita la actuación de los candidatos. Como nuestro alumnado es muy heterogéneo la elección de temas es muy amplia y variada.

Como ya vimos en el capítulo 3, (Alderson, 2000: 29) la lectura debe ser evaluada con una batería de técnicas centradas en el contenido. Los textos que tengan sentido para los lectores, que les interesen, que tengan relación con sus estudios, con sus aficiones e intereses o con su nivel intelectual pueden motivar una lectura más profunda que la de los textos tradicionales, anodinos y a veces sin mucho contenido. Los conocimientos de los candidatos influyen en la comprensión y por lo tanto se debe hacer un esfuerzo para permitir que este

conocimiento facilite sus resultados, en lugar de permitir que su ausencia inhiba sus actuaciones.

5.4. Conclusiones

En este breve relato de la historia de la evaluación de la EOI y la inquietud de los profesores por llegar a evaluar a nuestros alumnos de la forma más fiable, justa y válida posible, se ha visto la preocupación que existe entre los profesionales de las Escuelas por seguir mejorando nuestro modo de medir el grado de competencia lingüística de nuestros alumnos, tanto oficiales como libres.

A pesar de que hemos andado un largo camino, todavía existen problemas. Necesitamos seguir con las sesiones de práctica y estandarización de las destrezas de expresión oral y escrita, que al ser subjetivas son las más problemáticas. Además todavía no estamos totalmente satisfechos con el uso de alguna de las técnicas que estamos utilizando en las destrezas de comprensión de lectura y comprensión oral, por ejemplo, el test de rellenar huecos o *cloze* y el de *headings*, ya que no parecen tener mucha validez ni de constructo ni aparente.

El *cloze* parece que unas veces está evaluando conocimientos gramaticales y otras veces parece evaluar simplemente el conocimiento de vocabulario. Por otra parte, el *cloze* no parece congruente con los métodos de aprendizaje de los cursos y para algunos estudiantes es como si estuvieran resolviendo un rompecabezas que no les gusta y al cual no encuentran sentido.

Esta fue la causa principal de que se intente estudiar un nuevo método válido y fiable, que sirva para evaluar la comprensión lectora y que pueda sustituir al *cloze* que tiene tan poca validez aparente entre los profesores y alumnos de la EOI y que además no es fácil de elaborar. Con este fin, pues, se

comenzó este trabajo de investigación cuyos resultados se exponen en el siguiente capítulo.

Segunda Parte

INVESTIGACIÓN EMPÍRICA

Capítulo 6

PROCESO METODOLÓGICO

6.1. Objetivos del estudio

Con los problemas del capítulo anterior en mente parece ser que el C-test, que es fruto de la insatisfacción producida por el *cloze*, proporcionaría el remedio para la mayor parte de los problemas que teníamos planteados al evaluar la comprensión lectora. Debido a sus buenas propiedades, tales como el ser un test fácil de construir y corregir, más corto, que incluye más huecos, que es un instrumento válido y fiable, "performing everything that the *cloze* test promised" (Klein-Braley, 1984: 145), y siendo menos frustrante para los alumnos que el *cloze*, podría ser un buen sustituto del mismo.

De acuerdo con sus creadores, todos los tests creados según el principio de redundancia son herramientas para evaluar la competencia global de un idioma. Si eso es así, el C-test tendría que correlacionar de forma significativa con la batería de tests de la EOI.

Jafarpur (1995) investigó la frecuencia de supresión de palabras y su impacto en la dificultad del test y afirma que "changing the deletion start and /or ratio would produce different C-tests" (p. 200). Sin embargo, en este estudio el número de sujetos que hizo cada modelo del C-test (10 sujetos) era muy bajo. Además él cambió el punto de partida y la frecuencia de supresión de palabras simultáneamente y utilizó solamente un texto, lo cual va en contra del espíritu del C-test, cuya idea principal es utilizar varios textos para evitar el efecto que el conocimiento del tema pueda tener sobre los resultados del test.

Si el C-test se construye empezando en la segunda frase de cada texto y se suprimen las palabras de forma alternativa, comenzando con la 2ª ó 3ª palabra, entonces existirán solamente dos posibles modelos para cada texto y por lo tanto dos posibles modelos de C-test. La cuestión es saber si los dos modelos de cada texto y los dos modelos de C-test son equivalentes entre sí o no.

Un número de investigadores ha demostrado que las palabras gramaticales o de estructura son menos difíciles que las de contenido. Por lo tanto cuando los candidatos tienen que restaurar las palabras suprimidas, observaremos que el número de palabras funcionales restauradas es superior al número de palabras de léxico. Klein-Braley (1981; en Klein-Braley 1985) encontró que esto era cierto en los *clozes* y Alborn et al. (1959; en Klein-Braley 1985: 91) demostraron que cuando se suprimen palabras, la predicción de restauración de las que pertenecen a una misma clase, en general, está "inversely related to the size of that class". Por lo tanto, las palabras funcionales que pertenecen a una clase pequeña y cerrada son altamente predecibles.

También hemos visto que la frecuencia de las palabras que componen un texto es también importante para predecir la dificultad del mismo y aunque el número absoluto de palabras funcionales es pequeño, sin embargo, son palabras muy frecuentes en el uso de la de lengua. Klein-Braley (1985: 95) dice que "the difficulty of the individual deletions in a C-test is not only a function of their absolute frequency in the language but also of their embedding in the specific text". Además, palabras que estadísticamente son muy raras pueden tener una frecuencia inusualmente alta en el contexto de un texto específico, de forma que cuando un candidato trata de restaurar un C-test puede hacer uso de claves proporcionadas por la redundancia natural del texto.

El modelo de supresión de palabras cambia los términos de función y los de léxico suprimidos y también la cantidad de información que el texto mutilado proporciona. Por lo tanto, puede haber diferencias significativas entre los super-ítems que forman el C-test tanto a nivel léxico como funcional.

En este estudio, por consiguiente, intentamos contestar a las siguientes preguntas:

1. ¿Podría utilizarse el C-test como una alternativa al *cloze* en la batería de tareas del examen de comprensión lectora de la EOI?

2. ¿Serán equivalentes los dos subtests o super-ítems A y B creados con cada uno de los cuatro textos?

3. ¿Serán equivalentes los dos modelos de C-test construidos con los mismos textos pero empezando a suprimir palabras en distintos puntos?

4. ¿Cuál será la correlación entre el C-test y el *cloze* o el C-test y otros tests usados en los exámenes de la EOI?

5. ¿Habrá gran variación entre los términos léxicos y los funciónales restaurados?

Teniendo en cuenta las preguntas anteriores formulamos las siguientes hipótesis:

1. Existen diferencias significativas entre las medias de los términos funcionales y los léxicos recuperados.

2. Existen diferencias significativas entre los super-ítems creados tanto a nivel léxico como a nivel funcional.

3. La palabra con la que se empieza a mutilar el texto no afecta a los resultados finales de los examinandos, es decir, no hay diferencias significativas en la puntuación final de los C-tests independientemente del punto donde se empiece a mutilar el texto que luego se ha de restaurar.

4. No existen diferencias significativas entre los dos modelos de C-test. Es decir, que los dos modelos de C-test creados son equivalentes.

5. Las palabras pautadas se recuperan más fácilmente que las no pautadas.

6. Existen correlaciones significativas entre el C-test y el *cloze*.

7. Existe correlación entre el conjunto de pruebas de la EOI y el C-test.

8. Existe correlación significativa entre los tests de la comprensión lectora de la EOI y el C-test.

6.2. Método

6.2.1. Sujetos

Los participantes en esta investigación fueron 151 estudiantes de la EOI Jesús Maestro de Madrid que en marzo del 2007 estaban asistiendo a clases de tercer curso de la Escuela para examinarse en junio del Certificado que entonces se denominaba Elemental pero cuyo nivel, como ya hemos explicado en el capítulo cinco, equivalía al nivel del Examen de *First Certificate* de la Universidad de Cambridge. Igualmente, ahora el certificado equivalente al Ciclo Elemental se llama certificado de Nivel Intermedio y la batería de tests que lo componen es idéntica a la del certificado anterior.

Todos los alumnos que participaron en este trabajo de investigación hablaban español como primera lengua. Se suponía que todos ellos poseían un nivel de competencia de inglés intermedio y que eran alumnos motivados, ya que su asistencia a estos cursos no es obligatoria. Pertenecían a 8 clases

diferentes, que fueron elegidas al azar en un horario que iba desde las 8 de la mañana a las 10 de la noche.

Los examinandos formaron una muestra muy diversa compuesta por un 74,2% (n = 112) de alumnas y un 25,8% (n = 39) de alumnos. El número relativamente grande de alumnas refleja con precisión la proporción real que existe en la Escuela tanto entre los alumnos que asisten a clase como entre los que más tarde se examinan de Certificado.

El 59,6% de los examinandos eran menores de 30 años y el 40,4% eran mayores de esa edad. La mayoría de ellos (82,8%) tenían títulos universitarios y el 47% de los candidatos habían estudiado inglés durante más de 10 años. Solamente el 7,3% de ellos habían estudiado inglés durante un periodo menor a tres años. El 70,2% de la muestra usa el idioma inglés por diferentes razones, aunque no con mucha frecuencia. Asisten a clase con regularidad y la mayoría de ellos dicen leer y ver películas en inglés (81,5%) pero sólo de vez en cuando. También dicen que intentan pensar en inglés y no en su idioma materno cuando tienen que escribir o hablar en inglés.

Los examinandos no estaban familiarizados con la técnica del C-test, ya que era un método completamente nuevo para ellos, pero sí lo estaban con la batería de tests que se utilizaba en el certificado de Ciclo Elemental, puesto que prácticamente la totalidad de esas técnicas de examen se practicaban en clase.

El conocimiento de cualquier tema por parte de los alumnos de la EOI es muy amplio debido a la gran variedad de profesiones e intereses que poseen así como a las diferentes edades que tienen estos alumnos que pueden ir desde los 14 hasta más de 70 años.

6.2.2. Materiales

a) El C-test

El problema para elaborar un C-test que pudiera sustituir al *cloze* utilizado en la batería de métodos de la destreza de comprensión lectora, es la selección de los textos individuales que representaran adecuadamente el rango de habilidades que utilizan los alumnos de nivel intermedio para procesar el texto. Cuando los C-tests se utilizan como indicadores de la competencia general de un segundo idioma Klein-Braley (1997) y Connelly (1997) aconsejan seleccionar al azar textos de muestras auténticas que representen los tipos generales de lengua que el alumno se va a encontrar en la vida real. La dificultad es estimada intuitivamente por las personas que elaboran el test.

Al principio del proceso se eligieron varios textos, procedentes de fuentes relacionadas con la educación, sobre temas variados que pensábamos podrían interesar a nuestros alumnos y que al mismo tiempo fueran "neutral with respect to potential differences in their background knowledge" (Bachman 1990).

Se eligieron 12 textos cortos que trataban de temas familiares tales como " my struggle with cigarettes", "telephone selling", "acid rain", "greeting cards", "addicted to computers", "internet", etc. y se les entregó a un grupo de varios profesores, nativos y no nativos de la EOI de Jesús Maestro para que actuaran de "expert judgement" (Kobayashi 2002) o "standard setting" (Alderson et al. 1995) y eligieran el grupo de textos más adecuado con el que se pudiera elaborar el C-test.

En primer lugar, se rechazaron los textos que se consideraron menos adecuados, bien por ser excesivamente cortos o excesivamente largos. Después hubo que procesar el resto de los textos reunidos y para cada texto propuesto había que preguntarse si los alumnos de ese nivel serían capaces

de procesarlo o si por el contrario los alumnos de niveles más bajos no serían capaces de hacerlo, es decir, si el texto era apropiado y representativo para final de curso del nivel intermedio de la EOI. Se eligieron los pasajes que alcanzaron un mayor consenso entre los expertos.

Una vez puestos de acuerdo sobre los textos que iban a formar el C-test hubo que ordenarlos teniendo en cuenta la dificultad ascendente de los mismos, siguiendo los consejos de la extensa teoría de elaboración de C-tests. En este caso se tuvieron en cuenta principalmente los temas de los distintos textos, y se consideró que el más fácil sería el texto que trataba del "ejercicio físico", ya que era un tema muy conocido para ellos al estar incluido en prácticamente todos los libros de texto que se imparten en la EOI. El segundo más común sería el de "relax" y el menos común de todos ellos el de "vitaminas" que es un poco más específico aunque no utilice un vocabulario para especialistas. También se tuvo en cuenta el perfil de frecuencia léxica, que de acuerdo con Read (2000) influye en la dificultad del texto. En el primer texto, "physical exercise", el 80% de las palabras de léxico que se tienen que recuperar tienen una frecuencia superior a 6.000 según el BNC, mientras que en el último, "vitamins are vital, solamente el 54% de las palabras de léxico que se tienen que recuperar tienen una frecuencia superior a 6.000.

De acuerdo con las directrices establecidas por Klein-Braley (1997) se construyeron dos modelos de C-test (C-test A y C-test B) formados por los cuatro mismos textos y con 100 palabras mutiladas cada uno. Cada texto contenía 25 palabras mutiladas pero el punto en el que se comenzaba a mutilar las palabras era diferente.

A los cuatro textos que forman el C-test se les llama también subtests o super-ítems. Nosotros utilizaremos indistintamente uno u otro término a lo largo de la investigación.

Modelo A. Consistía en el siguiente diseño de supresión de palabras:

- Texto 1: "*Physical exercise"* se empezó a suprimir la segunda palabra de la segunda frase.
- Texto 2: "*Relax and live*" se empezó a suprimir la segunda palabra de la segunda frase.
- Texto 3: "*Alternative sources of energy*" se empezó a suprimir la tercera palabra de la segunda frase.
- Texto 4: "*Vitamins are vital*" se empezó a suprimir la tercera palabra de la segunda frase.

Modelo B. Consistía en el siguiente diseño de supresión de palabras:

- Texto 1: "*Physical exercise*" se empezó a suprimir la tercera palabra de la segunda frase.
- Texto 2: "*Relax and live"* se empezó a suprimir la tercera palabra de la segunda frase.
- Texto 3: "*Alternative sources of energy*" se empezó a suprimir la segunda palabra de la segunda frase.
- Texto 4: "*Vitamins are vital*" se empezó a suprimir la segunda palabra de la segunda frase.

En cada texto, independientemente de la dificultad de las palabras, se suprimió la segunda mitad de una de cada dos palabras. En ambos modelos de C-test se pautaron los huecos en los textos o super-ítems "*Physical exercise*" y "*Alternative Sources of Energy*", es decir, se sustituyó cada letra suprimida por un guión. Por ejemplo, la palabra "from" quedaría en el texto como "fr- -". En cambio en los otros dos textos de los cuatro que forman el C-test: "*Relax and*

Live" y "*Vitamins are Vital*", los huecos no se pautaron sino que las letras suprimidas se reemplazaron por un único guión, por ejemplo "fr__", hasta que se suprimieron 25 palabras de cada texto. El resto de cada texto se dejó intacto hasta el final. Si las palabras tenían un número impar de letras, siguiendo la teoría de la construcción del C-test, se suprimía la segunda mitad más una letra; ej." tires" se convertía en "ti - - -" o "ti___", según si se encontraba en un texto pautado o no. Estas reglas de supresión de letras no se aplican si las palabras constan de una sola letra, son números, fechas, acrónimos o nombres propios. En este caso las palabras no se mutilan.

Puesto que esta técnica de examen era totalmente desconocida para los alumnos, se dieron instrucciones escritas al principio del C-test y se les proporcionó un ejemplo también por escrito. El día del examen se explicó en clase oralmente en qué consistía el test y se admitieron preguntas para aclarar cualquier duda que tuvieran los alumnos con respecto al modelo de test que tenían que hacer.

Los dos modelos de C-test se reproducen en el apéndice 1. Generalmente, los huecos dentro de un texto dado del C-test son localmente dependientes en un grado significativo. De esta forma "C-test texts have to be construed as super-items or testlets, with item values corresponding to the number of blanks filled in correctly; that is, each text represents a politomous item" (Eckes and Grotjahn 2006: 305). El C-test que hemos construido consta de cuatro textos o super-ítems, y cada uno de ellos puede obtener valores de entre 0 y 25.

b)- La batería de tests de la EOI que consistía en los siguientes tests:

- Un test de comprensión auditiva de preguntas cortas.
- Un test de comprensión auditiva de elección múltiple.
- Un test de comprensión auditiva para establecer correspondencia entre títulos y extractos de distintos textos auditivos.

- Un test de comprensión lectora de elección múltiple.
- Una tarea de comprensión lectora que consiste en casar títulos con extractos de lectura y a la que hemos denominado "Headings".
- Un *cloze* de banco de palabras, es decir, un test de comprensión lectora de rellenar huecos, proporcionando un banco de palabras con las respuestas y varios distractores.
- 2 tareas de expresión escrita.

Las características de todos los tests de la EOI han sido descritas en el capítulo 5 (la Evaluación en la Escuela de Idiomas) y se ha dicho que están pilotadas y revisadas por un equipo experto que trabaja para la comunidad de Madrid. Desgraciadamente, y por razones administrativas, la prueba de expresión oral no se pudo realizar junto con el resto de los exámenes.

c)- *El test del tutor*

Consiste en una tarea similar a la del C-test a la que se le ha llamado "test del tutor" o "test de control" (apéndice 2) y que consta de 18 frases en las que se ha mutilado una palabra que los candidatos tienen que restablecer. Las frases se seleccionaron para comprobar la homogeneidad de competencia lingüística de los candidatos y así poder hacer comparaciones entre los resultados obtenidos por los candidatos que hicieron el C-test A y los que hicieron el C-test B. Con este procedimiento se evitó tener que repetir el mismo examen al cabo de un tiempo con los problemas que ello acarrea debido al posible aprendizaje, memoria de los textos, etc. Si los grupos son homogéneos podremos considerarlos como si se tratara de un mismo grupo a la hora de evaluar los resultados y considerar la fiabilidad y la validez de los mismos.

d)- *El cuestionario*

Se elaboró un cuestionario (apéndice 3) para recoger datos sobre los candidatos además de sus actitudes, sus sentimientos, sus opiniones sobre lo que este test medía y finalmente, sus reacciones sobre el C-test, y de este modo poder analizar la validez aparente del mismo.

6.2.3. Procedimiento

Los dos modelos de C-test y la batería de exámenes de la EOI se administraron a los alumnos con un intervalo de una semana en las dos primeras semanas de marzo de 2007.

La primera semana los alumnos se examinaron de los tests de la EOI bajo la supervisión de sus propios profesores. Los tests objetivos, como fueron los tests de comprensión lectora y los de comprensión auditiva, fueron corregidos por sus propios profesores usando la clave de corrección que proporcionó el Departamento de Inglés de la Escuela Oficial de Idiomas de Jesús Maestro de Madrid. Sin embargo, los dos tests de expresión escrita, al ser pruebas subjetivas, fueron corregidos por dos expertos profesores de inglés que utilizaron las tablas analíticas de calificación diseñadas también por el departamento de inglés y que se utilizan en todos los exámenes de nivel intermedio que se realizan en dicha Escuela. Los tests de la EOI se administraron como criterio externo para poder estudiar la validez concurrente y la fiabilidad de los dos modelos de C-test que se habían elaborado.

La segunda semana se administraron los dos modelos de C-test, el test del Tutor, y el Cuestionario. A los profesores encargados de cada clase se les proporcionó instrucciones detalladas sobre la prueba y su administración. También se indicó a los alumnos el procedimiento a seguir para realizar el test.

Cuando los alumnos llegaron a clase eligieron libremente el lugar donde sentarse. Una vez leídas las instrucciones se contestaron las dudas o preguntas que quisieron plantear sobre los exámenes y el cuestionario y se distribuyeron los dos modelos de C-test totalmente al azar. También se distribuyeron el test del Tutor y el Cuestionario que eran iguales para todos.

El Test del Tutor, como ya se ha dicho, se usó como un elemento común externo para estudiar la homogeneidad de los dos grupos. Se les dio instrucciones para que contestaran en las mismas hojas de los tests y se les concedió un tiempo de 30 minutos para los dos tests más el tiempo que necesitaran para contestar al Cuestionario.

Tanto los dos modelos de C-test como el test del Tutor se corrigieron siguiendo el método de la respuesta exacta, lo que, por definición, significa que únicamente se consideraba respuesta correcta la palabra exacta que se había mutilado en el pasaje original. No se aceptaron ni errores de ortografía, ni sinónimos, ni otras palabras que aunque no fueran la palabra exacta del texto original podrían haber tenido sentido en la frase. La razón para usar este método fue evitar introducir manipulación alguna por parte del investigador, ya que según afirma Brown (1993: 98): “research indicates high correlations between exact-answer scoring results and other scoring procedures”. Las respuestas se cuantificaron de acuerdo con los términos léxicos y funcionales.

Se pidió a los alumnos que escribieran sus nombres tanto en los tests como en el Cuestionario. Como recomienda Dörnyei (2003), se les prometió confidencialidad y que los nombres serían eliminados en el trabajo de investigación, ya que iban a ser asociados con un número, con el objetivo de facilitar la aplicación de algunos estadísticos en los tests de la EOI, los C-tests, el test del Tutor y el Cuestionario.

Lo ideal habría sido que todos los candidatos hubieran hecho los dos C-tests (C-test A y C-test B), facilitando así la comparación de los resultados. Sin embargo, este procedimiento habría debilitado la validez de la investigación, que se habría visto socavada si todos los candidatos hubieran

tenido acceso a los mismos textos más de una vez, ya que podían haber recordado los textos o haberse sentido más cómodos con el formato de examen, y por lo tanto su rendimiento habría sido mejor en la segunda administración del test. Este problema se superó aumentando el tamaño de la muestra y asegurándose de que los grupos tenían el mismo nivel de competencia de la lengua. El test del Tutor tuvo la función de demostrar que el grupo A y el grupo B eran grupos homogéneos.

6.2.4. Herramientas para el análisis estadístico de los datos

Como dice Bachman (2004: 33) a menudo nos encontramos con gran cantidad de datos que provienen de diferentes individuos y de diferentes pruebas que tienen que ser resumidos para entenderlos e interpretarlos. Los métodos estadísticos nos proporcionan las herramientas para realizar esa interpretación así como para poder hacer inferencias y generalizaciones sobre el comportamiento de grupos más numerosos de individuos. Para analizar e interpretar la gran cantidad de datos que se obtuvieron en los tests de este estudio se utilizó el programa estadístico *Statistical Package for Social Sciences* (SPSS 15.0).

Gracias a este programa se han estudiado las correlaciones entre todos los tests y subtests de los que consta el estudio, la fiabilidad de los mismos y el peso específico de cada prueba dentro de cada test global. Así mismo, se han realizado estudios para determinar el grado de facilidad de recuperación de los términos mutilados concentrándonos especialmente en los términos léxicos.

Por último, se estudiaron las respuestas del cuestionario para determinar la validez aparente del C-test y analizar si existía alguna relación entre los resultados del C-test y algunas características de los candidatos, como pueden ser la edad, el sexo o los hábitos que tienen con respecto al uso de la lengua.

Capítulo 7

ANÁLISIS DE LA RECUPERACIÓN DE LOS TÉRMINOS LÉXICOS

7.1. Introducción

Queremos empezar por el análisis de los términos que constituyen el C-test antes de entrar en si hay diferencias o correlaciones entre los distintos tipos de tests.

Al diseñar los dos modelos de C-tests A y B decidimos estudiar si facilitaría la recuperación de las palabras el proporcionar a los alumnos el número de letras que faltaban de una determinada palabra. Así pues, al construir los C-tests de cuatro textos distintos, extraídos del material paralelo al que se utiliza en ese nivel, el texto primero llamado PHYSICAL EXERCISE y el tercero llamado ALTERNATIVE SOURCES OF ENERGY son pautados tanto en el C-test A como en el C-test B, es decir, indicamos a los alumnos el número de letras que tienen que añadir para completar la palabra. Los otros dos textos, el segundo titulado RELAX AND LIVE y el cuarto titulado VITAMINS ARE VITAL están sin pautar, es decir, los alumnos no saben si el número de letras suprimidas es el mismo que el número de letras que permanecen en el hueco o si se ha suprimido una más de las que permanecen en la palabra mutilada.

7.2. Sustantivos

Los resultados de la tabla 7.1 nos indican que hay una gran diferencia entre la proporción de ítems pautados y no pautados recuperados. De los 17 ítems que han sido recuperados por un número de alumnos ≥ 60 en cada uno de los dos grupos[1], 76,5% corresponden a ítems pautados y el 23,5% a ítems no pautados. En cuanto a los 12 ítems que peor se han recuperado, ya que sólo lo han recuperado un número de alumnos ≤ 40, la mayoría corresponden a ítems no pautados, en concreto un 66,7%. Esto nos demuestra que los ítems

[1] El grupo A se componía de 77 alumnos y el Grupo B de 74 alumnos.

que son recuperados por un porcentaje muy alto de los alumnos corresponden en su inmensa mayoría a ítems pautados y los que son recuperados por un porcentaje muy bajo de alumnos (≤ 40) corresponden mayoritariamente a ítems no pautados.

Los resultados expuestos anteriormente y reflejados en la tabla 7.1 nos indican que los textos sin pautar tienen una dificultad añadida a la hora de recuperar las palabras mutiladas, ya que los alumnos no saben exactamente el número de letras que se han suprimido cuando se mutiló la palabra, y a veces cometen errores de ortografía, morfología o de gramática.

Tabla 7.1

	Nº Sust. Recup.	Ítems Pautados	% Recuper.	Ítems no Pautados	% Recuper.
Recuper. ≥ 60	17	13	76,5	4	23,5
Recuper. ≤ 40	12	4	33,3	8	66,7

Como ya observaron Kontra y Kormos (2006:129), los alumnos emplean muy pocas estrategias morfológicas para recuperar las palabras, y a veces se limitan a contar el número de letras que tienen que reponer en lugar de analizar y encontrar pistas en el contexto sintáctico en el que se encuentra la palabra mutilada. Por lo tanto, cuando estas pistas no existen, por no estar el texto pautado, cometen más errores.

Veamos las siguientes tablas en las que se presenta cada uno de los términos que se tenían que recuperar, la frecuencia de estos términos en el British National Corpus y las frecuencias absolutas y relativas de las respuestas correctas e incorrectas.

SUSTANTIVOS ≥ 60

Tabla 7.2

C-TEST A					
Sustantivos ≥ 60	BNC	Correct.	Incor.	% Cor.	% Incor.
People	121663	76	1	98,7	1,3
Year	73046	75	2	97,4	2,6
World	57430	60	17	77,9	22,1
Body	24590	60	17	77,9	22,1
Rate	18523	70	7	90,9	9,1
Friend	14377	65	12	84,4	15,6
Energy	12091	72	5	93,5	6,5
Weight	8189	72	5	93,5	6,5
Exercise	4852	74	3	96,1	3,9
Illness	3214	70	7	90,9	9,1

Tabla 7.3

C-TEST B					
Sustantivos ≥ 60	BNC	Correct.	Incor.	% Cor.	% Incor.
Time	151834	73	1	98,6	1,4
Way	95412	60	14	81,1	18,9
Body	24590	66	8	89,2	10,8
Population	13005	62	12	83,8	16,2
Exercise	4852	73	1	98,6	1,4
Individual	3439	60	14	81,1	18,9
Substances	1330	70	4	94,6	5,4

Aunque como ya se ha comprobado, la mayoría de los sustantivos recuperados corresponden a huecos pautados, sin embargo, existen excepciones de sustantivos que a pesar de pertenecer a los textos que no

están pautados se han recuperado fácilmente. Estos son: *illness*, *friend*, *way* y *substances*. Las explicaciones que podemos dar para ello son varias.

a. *Friend* – Este es un sustantivo muy frecuente en el vocabulario de los alumnos. Además aparece en el texto en una frase como nombre opuesto a *enemy.* "stress can be a fri___ or an enemy", lo cual ayuda a su recuperación (Carnine et al, 1984 y Neuner, 1992).

b. *Way* – Además de que es una palabra muy frecuente en el BNC (95.412) y de que ya forma parte del vocabulario activo de los alumnos de este nivel por aparecer muy pronto en sus libros, en el texto figura como parte de la expresión "way of life". Esta expresión tiene una frecuencia en el BNC de1.036, relativamente muy alta para ser un grupo de palabras. Ello se debe a que forman lo que Nattinger y DeCarrico (1992) llaman "lexical frase" y Biskup (1992) colocación típica, que es un grupo de palabras que operan como una unidad, con una función particular en el discurso hablado y escrito.

c. *Substances* – Su frecuencia en el BNC no es muy alta (1.130). Sin embargo, la palabra es muy semejante al español y además aparece en el texto justamente en la línea anterior, con lo cual es muy fácil encontrar la correlación entre la frase del texto: "certain substances were essential in the diet to prevent or cure some diseases" y la frase en la que aparece la palabra mutilada: "these subst___ are known as vitamins".

d. *Illness* – A pesar de no ser una palabra muy frecuente en el BNC (3.214), sin embargo, las palabras *ill* e *illness* sí forman parte del vocabulario de los alumnos. También aparece en el texto este sustantivo en plural *illnesses*. Por otra parte se dan varios ejemplos de enfermedades tales como "heart attacks" o "alcoholism" en la frase siguiente.

Merece la pena comentar también las palabras *Individual* y *exercise* cuya frecuencia cuya frecuencia en el BNC no es muy alta: 3.439 y 4.852

respectivamente. Sin embargo, consideramos que han sido recuperadas fácilmente por un número muy alto de alumnos por varios motivos:

1. Son palabras similares al español (ver Abbott, 2007).
2. Son palabras pautadas.
3. Se les proporciona bastante información de la palabra, ya que al no ser palabras muy cortas disponen de cuatro y cinco letras respectivamente para recuperar la palabra completa.

Por otra parte, hay que tener en cuenta que aunque en el BNC la palabra *exercise* como sustantivo no sea excesivamente frecuente, la frecuencia general de esa palabra como verbo y como sustantivo es mucho mayor (8.590). Además, para un alumno *exercise* es una de las palabras que más utilizan, leen y oyen a diario en clase. Podemos afirmar que en el ámbito en el que ellos practican el inglés, la palabra *exercise* tiene una frecuencia máxima.

De acuerdo con Alderson (2007a: 405), la frecuencia de una palabra en el uso de la lengua de una persona depende de su experiencia con las palabras, la cual puede ser muy diferente de la de otras personas. El contexto en el que se usa un término léxico es tan importante como la frecuencia (Schmitt, 2000: 16). Así la palabra *exercise*, que estamos analizando, no es particularmente frecuente en inglés general pero es indispensable en la clase, por lo que esta palabra será más fácilmente recuperada que cualquier otra que pueda tener una frecuencia mayor en el BNC pero que no sea de uso tan frecuente entre los alumnos

SUSTANTIVOS ≤ 40

Tabla 7.4

C-TEST A					
Sustantivos ≤ 40	BNC	Correct.	Incor.	% Cor.	% Incor.
Food	18639	19	58	24,7	75,3
Growth	12787	18	59	23,4	76,6
Pressure	11511	16	61	20,8	79,2
Functions	4802	13	64	16,9	83,1
Maintenance	3955	10	67	13,0	87,0
Amounts	2402	38	39	49,4	50,6
Deficiency	670	9	68	11,7	88,3

Tabla 7.5

C-TEST B					
Sustantivos ≤ 40	BNC	Correct.	Incor.	% Cor.	% Incor.
Rest	13208	24	50	32,4	67,6
Scientists	3485	35	39	47,3	52,7
Signals	1686	22	52	29,7	70,3
Vitamin	822	26	48	35,1	64,9
Tiredness	283	16	58	21,6	78,4

Como ya se ha dicho, el 66,7% de los sustantivos que han sido recuperados por un número de alumnos ≤ 40 corresponden a huecos sin pautar. Las excepciones son*: tiredness*, *amounts, scientists* y *rest*, que a pesar de ser pautadas no se han recuperado fácilmente.

1. *Tiredness* – Es una palabra muy poco frecuente, ya que solamente tiene una frecuencia de 283 en el BNC y definitivamente podemos decir que

no es una palabra que forme parte ni del vocabulario activo ni del pasivo de alumnos de nivel intermedio.

2. *Amounts* – Al considerar en la corrección de los tests como respuesta correcta solamente la palabra exacta a la del texto original, no se ha aceptado *amount* que muchos alumnos han dado y también es posible en la expresión "... new ways of generating large amount(s) of energy". Lo mismo ha ocurrido con las palabras *functions* en la frase "... the normal function(s) of the body" y *signals* en la frase "… if you do not notice the warning signal(s) ... ".

3. *Scientists* – Con esta palabra nos encontramos con dos problemas, el de la flexión de plural y el de la formación de palabras que es un problema añadido. La lengua española utiliza la misma palabra "científico" como adjetivo y como sustantivo y los alumnos tienden a hacer lo mismo en inglés asimilándolo al español y usando *scientific* en lugar de *scientist* para el sustantivo de personas. Otro problema puede haber sido el confundir la clase de palabra por lo que han usado el adjetivo en lugar del sustantivo (Odlin y Natalico, 1982).

4. Por último, comentaremos la palabra *rest*, que a pesar de ser pautada, tener una frecuencia intermedia (13.208) y ser una palabra corta, que de acuerdo con toda la literatura existente son más fáciles de aprender, es recuperada solamente por 24 de los 74 alumnos que realizaron ese modelo de test.

Analizando las expresiones del BNC vemos que en el 90% de los ejemplos más comunes usan *rest* con el significado de "resto" y en muy pocos casos tiene el significado de "descanso", que es el que tiene en el texto.

Ejemplos:

- The rest of the day
- The rest of the cycle
- The rest of his life

- The rest of the production process
- The rest of mankind
- We flew the rest of the way home in silence

También observamos que muy frecuentemente *rest* aparece formando parte de la expresión "the rest of". De hecho de las 13.208 veces que aparece *rest* en el BNC, 8.788 corresponden a la expresión "rest of" y 8.683 aparece formando parte del grupo de palabras "the rest of". La presencia de *rest* con el significado de "descanso" y conjuntamente con el verbo *take*, (take __ rest), es muy escasa e incluso cuando aparece con este verbo, el significado sigue siendo el de "resto". Ej. "Take the rest of the bread". Entre todos los ejemplos registrados en el BNC la expresión "take a rest" aparece solamente una vez y no existe ningún ejemplo de "take some rest" donde *rest* sea un nombre incontable.

A continuación vamos a analizar otros sustantivos cuyos resultados de recuperación pueden parecer sorprendentes, tales como: *growth, deficiency, pressure y maintenance*. Observamos que todos ellos son términos derivados que utilizan sufijos para formar nuevas palabras. La formación de palabras por lo tanto, que supone cierta dificultad para el aprendizaje de nuevo léxico, puede ser también un problema añadido para la recuperación del mismo en el C-test.

1.- *Growth* - Si prestamos atención a la frecuencia de estas palabras en el BNC, comprobamos que la más frecuente es *growth* con una frecuencia de 12.787. Esto quiere decir que de acuerdo con el BNC *growth* es una palabra mucho más frecuente en el idioma inglés que el verbo *grow* cuya frecuencia de acuerdo también con el BNC es de 5.335. Sin embargo, en el vocabulario pasivo, e incluso me atrevería a decir que también en el activo, de estudiantes de inglés como lengua extranjera de nivel intermedio, que fue con los que realizamos la investigación, la frecuencia del verbo *grow* es mucho mayor que la del sustantivo *growth*.

La introducción a los estudiantes del verbo *grow* ocurre con mucha mayor anterioridad que el sustantivo *growth*, lo que significa que para ellos, en

el discurso oral y escrito que utilizan, el verbo es mucho más frecuente que el sustantivo.

Hemos analizado ambas palabras en algunos ejemplos del BNC y comprobamos que los temas y frases donde aparecen presentan también un grado de dificultad muy distinto.

Ejemplos:

- The plants will grow…
- All children want to grow up…
- I'm not going to grow old…
- The annual turnover growth…
- Early growth in the East…
- The current expansion and urban growth in the Southern population…

El verbo *grow* o su variante *grow up* son bastante frecuentes en el vocabulario que utilizan los alumnos en este nivel. Sin embargo, el nombre *growth* no lo es. Por otra parte, *growth* tiende a colocarse con palabras como *turnover* o *urban* que añaden dificultad a su comprensión.

Esta información nos lleva a concluir que aunque la frecuencia de una palabra en el idioma es muy importante para su aprendizaje y su uso, sin embargo, es todavía más importante la frecuencia relativa de esa palabra en el vocabulario de los estudiantes de inglés como lengua extranjera a un cierto nivel.

2.- *Deficiency* - la palabra *deficiency* no está pautada y es muy poco frecuente tanto en el BNC como en el vocabulario de los alumnos de este nivel. Sin embargo, los que han seguido el sentido del texto y no han podido recuperar esta palabra, la han sustituido por *deficit* que es mucho más frecuente (2.346). Esta palabra al ser más corta que la mutilada no cumple una de las reglas del C-test que dice que el número de letras proporcionadas tiene

que ser igual o menor al número de letras que tienen que ser añadidas para recuperar la palabra completa (defic_____). Sin embargo, los candidatos sí que han mantenido el sentido y la clase de palabra que se les pedía.

3.- *Pressure* – Esta palabra tampoco está pautada (pres_____) por lo que los candidatos saben que las letras que faltan para completar la palabra pueden ser tres o cuatro. Algunos de los errores han sido de ortografía al dar como respuesta *pressure* con una sola "s", que por el número de letras que faltaban podía haber sido posible. Otros alumnos forman una palabra nueva *pression*. Para la formación de esta palabra toman como modelo *depress* cuyo sustantivo es *depression*, usada como una enfermedad, una depresión del terreno o una depresión económica.

Ejemplos:

- The economic depression
- The depression after the war
- He's suffering from depression
- They depress me

Aplicando la misma regla y si del verbo *depress* se forma el nombre *depression*, puede parecer lógico que del verbo *press* se pueda formar el nombre *pression*, que por otra parte es una palabra afín a la palabra española presión. Luego, por una parte, han generalizado una regla de formación de palabras y por otra la han asimilado a la formación de palabras del español. Este resultado está de acuerdo con la teoría de Singleton y Little (1991) que dice que los alumnos cuando no conocen una palabra tienden a crear una palabra posible pero incorrecta o plausible pero no existente.

Nesselhauf (2003) estudió la influencia de la primera lengua en los distintos tipos de errores que los estudiantes cometían obteniendo resultados que indicaban que alrededor de la mitad de los errores que cometían los alumnos se debían a la influencia de su primera lengua. Esto nos lleva a la conclusión de que la tendencia de las últimas décadas de minimizar la

influencia de la primera lengua en el aprendizaje de una segunda lengua parece equivocado.

4.- *Maintenance* – Este sustantivo no es muy frecuente en el BNC (5.955) y aún menos frecuente en el vocabulario activo de los alumnos. El sufijo utilizado en esta palabra para formar el sustantivo no es tampoco de los más utilizados en este nivel de inglés.

Por último, queremos comentar la palabra *vitamin* que no es una palabra muy frecuente en el BNC (822), pero que a simple vista no parece una palabra que pudiera plantear ninguna dificultad para los hablantes de español, sobre todo si tenemos en cuenta que está incluida en un texto cuyo título es VITAMINS ARE VITAL. Sin embargo, en este texto aparece colocada delante del nombre *deficiency* por lo que está actuando como adjetivo. Por ello, muchos alumnos han dado como respuesta el adjetivo *vital* y los que han pensado en la palabra *vitamin* han dado su forma plural *vitamins*. En este caso, la gramática inglesa ha jugado en su contra, ya que todavía en niveles intermedios aplican las reglas de concordancia de la gramática española al inglés y suelen utilizar los adjetivos en plural. Así pues, para la mayoría de ellos "deficiencia de vitaminas" se traduciría como "vitamins deficiency".

Estos resultados están en la linea de los obtenidos por Kontra y Kormos que encontraron que dentro de las estrategias metacognitivas muy pocos candidatos aplican estrategias selectivas. De acuerdo con su investigación, cuando los alumnos relataron como fueron capaces de restaurar la palabra mutilada vieron que muy pocos examinandos se daban cuenta de que la palabra mutilada podría encontrarse sin mutilar en el mismo texto.

> Only a few of the students expressed awareness that occasionally a truncated word might occur in the text in a non-truncated form, or one item in a list might give a clue regarding the other items in the same list. (Kontra y Kormos, 2006:133)

7.3. Adjetivos

La presencia de los adjetivos, como se espera en este tipo de textos, es inferior a la de los sustantivos.

ADJETIVOS ≥ 60

Tabla 7.6

C-TEST A					
Adjetivos ≥ 60	BNC	Correct.	Incor.	% Cor.	% Incor.
Same	61173	69	8	89,6	10,4
Better	20774	73	4	94,8	5,2

Tabla 7.7

C-TEST B					
Adjetivos ≥ 60	BNC	Correct.	Incor.	% Cor.	% Incor.
Normal	12176	65	9	87,8	12,2
Optimistic	1220	61	13	82,4	17,6

Aunque solamente cuatro adjetivos han sido recuperados por un número de alumnos ≥ 60, por lo que estadísticamente no pueden ser considerados, sin embargo, podemos constatar que el 75% de ellos corresponden a palabras pautadas.

El adjetivo no pautado que se ha recuperado es *normal* que es igual que su homónimo español. Esto explica que haya sido recuperado por 65 alumnos

de los 74 que hicieron el C-test B. El adjetivo *optimistic* es muy poco frecuente en el BNC. Sin embargo, tiene la misma raíz que en español y es uno de los adjetivos que se aprenden relativamente pronto cuando se enseña a describir a las personas en cursos de nivel básico, con lo cual la frecuencia de este adjetivo en el vocabulario activo de los alumnos es alta.

ADJETIVOS ≤ 40

Tabla 7.8

C-TEST A					
Adjetivos ≤ 40	BNC	Correct.	Incor.	% Cor.	% Incor.
General	22978	31	46	40,3	59,7
Warning	344	27	50	35,1	64,9

Tabla 7.9

C-TEST B					
Adjetivos ≤ 40	BNC	Correct.	Incor.	% Cor.	% Incor.
New	113654	34	40	45,9	54,1
Good	76617	19	55	25,7	74,3
Large	34236	30	44	59,5	40,5
Western	9595	22	52	29,7	70,3
Present	7213	16	58	21,6	78,4
Balanced	958	21	53	28,4	71,6

Llaman la atención algunos adjetivos cuya frecuencia en el BNC es alta y, sin embargo, han tenido un número de aciertos muy bajo. Por ejemplo, *new*, *good*, *large* y *present*.

1. Tanto las palabras *new* como *good* son muy frecuentes y conocidas para este nivel de inglés. Sin embargo, ambas son muy cortas con lo que la información o letras proporcionadas en el ejercicio son muy escasas. En *new* se les daba simplemente la "n" (n - -) y en *good* se daba "go" (go___). El número de palabras que empezando con "n" tienen dos letras más es muy amplio en inglés y los candidatos han dado varias: *now, not, nor, non*, etc. Lo mismo ha ocurrido con "go___" que además al no estar pautada esta palabra podía constar de dos o tres letras más. Ej. *goes, going*, etc.

2. *Large* (la - - -). Además de presentar el problema anterior, *large*, que en el BNC tiene una frecuencia de 34.236, no es una palabra que ellos utilicen a menudo. En su vocabulario es mucho más frecuente el uso de *big*, aunque en el BNC esta palabra tenga una frecuencia menor (24.861). Para un alumno de nivel intermedio español utilizar *big* es mucho más natural que utilizar *large* y así veremos que expresiones tales como:

    - a big amount
    - a big proportion of the women
    - is this a big number or a small number?
    - it was a big desk
    - he is a big man
    - it was a big firm

    son mucho más comunes en sus expresiones habladas y escritas que:

    - a large amount
    - a large amount of women
    - is this a large number or a small number?
    - it was a large desk
    - he is a large man
    - it was a large firm

Con lo cual constatamos de nuevo que la facilidad de recuperación de un término léxico depende efectivamente de la frecuencia del mismo, pero no de su frecuencia total dentro del idioma sino de la frecuencia y presencia que ese término tenga en el vocabulario activo de los alumnos de un determinado nivel.

3. Otro ejemplo lo tenemos también en el adjetivo *present* con el significado de "actual" en español. En este caso, los alumnos tienden a usar la palabra *actual* en inglés, ya que es igual que la española y por lo tanto se identifican con ella, la aprenden y la utilizan mucho más frecuentemente.

En el BNC la expresión "present rate" (47) es más frecuente que "actual rate" (32), y la palabra *present* (7.213) es más frecuente que *actual* (6.777). Sin embargo, en el vocabulario de un español que hable inglés las frecuencias se invertirán debido a la similitud de la palabra *actual* con la de su propio idioma. Así pues, vemos que la influencia del idioma materno de un alumno de inglés es tremendamente importante a la hora de recuperar el léxico del test y puede influir más que la frecuencia de uso que esa palabra pueda tener entre los hablantes nativos de Inglés.

También ha sido una sorpresa que la palabra *general* que es una palabra que tiene la misma forma en el idioma español y cuya frecuencia es relativamente alta (22.978) haya sido recuperada solamente por 31 alumnos. Entre las respuestas más comunes que han dado se encuentran: *genetic*, *genuine*, *gentle* y *generic*. Como puede observarse todas estos términos son adjetivos lo mismo que la palabra *general*, sin embargo, estas palabras no han sido consideradas aunque hubieran sido posibles y tuvieran sentido con el resto del texto, ya que sólo se han aceptado como correctas las palabras exactas.

7.4. Verbos

VERBOS ≥ 60

Tabla 7.10

C-TEST A					
Verbo ≥ 60	BNC	Correct.	Incor.	% Cor.	% Incor.
Developed	7668	60	17	77,9	22,1
Suffer	1355	69	8	89,6	10,4

Tabla 7.11

C-TEST B					
Vebos ≥ 60	BNC	Correct.	Incor.	% Cor.	% Incor.
Feel	12307	61	13	82,4	17,6
Damage	766	63	11	85,1	14,9

Solamente 4 verbos han sido recuperados por un número de alumnos igual o mayor de 60. Sin embargo, tienen en común que todos ellos corresponden a textos de palabras pautadas y se encuentran en la banda baja de frecuencias del BNC. Esto puede considerarse como un dato más, aunque estadísticamente no sea representativo ni digno de tener en cuenta por el escaso número de palabras, que indica que el pautar los huecos ayuda a la recuperación de las palabras.

El verbo *damage* tiene una frecuencia muy pequeña como verbo (766), sin embargo, la frecuencia de esta palabra si consideramos su función tanto como verbo como sustantivo es considerablemente mayor (8.301), lo que hace que esta palabra sea bastante más conocida para los alumnos (al menos como sustantivo) de lo que a primera vista cabría esperar.

Una vez más vemos que la frecuencia de las palabras en el idioma es importante para predecir la dificultad de aprendizaje de las mismas tanto de forma activa como pasiva. No obstante, el grado de influencia es muy complejo de determinar, ya que depende de un número de variables muy amplio que también hay que considerar en cada momento.

VERBOS ≤ 40

Tabla 7.12

C-TEST A					
Verbos ≤ 40	BNC	Correct.	Incor.	% Cor.	% Incor.
Provide	16292	22	55	28,6	71,4
Buy	9384	7	70	9,1	90,9
Continues	3956	10	67	13.0	87,0
Require	3052	29	48	37,7	62,3
Notice	1984	29	48	19,2	31,8
Increasing	1781	34	43	44,2	55,8
Warn	748	9	68	11,7	88,3
Generating	262	12	65	15,6	84,4
Tires	34	14	63	18,2	81,8

Tabla 7.13

C-TEST B					
Verbos ≤ 40	BNC	Correct.	Incor.	% Cor.	% Incor.
Look	21429	38	36	51,4	48,6
Makes	16332	40	34	54,1	45,9
Benefit	2343	31	43	41,9	58,1
Fear	1163	14	60	18,9	81,1
Burns	202	27	47	36,5	63,5

Lo primero que observamos es que hay muchos más verbos que han sido recuperados por un número de alumnos ≤ 40 que los que han sido recuperados por un número ≥ 60. Esto es lo mismo que ocurría con los adjetivos pero lo contrario que ocurría con los sustantivos. Los resultados pues están de acuerdo con la teoría de Ellis y Beaton (1993) que dice que un sustantivo es más fácil de aprender y recuperar que un adjetivo o un verbo, ya que es más fácil reproducir una imagen visual de su significado en nuestro cerebro. Nuestro cerebro reproduce y asocia más fácilmente la imagen de "pan" o "silla", por ejemplo, con sus respectivos significados, que la imagen del verbo "ir" o "beneficiarse" con sus posibles significados. Esto hace que un sustantivo sea retenido y recuperado más fácilmente que un adjetivo o un verbo.

También se observa que las flexiones y los sufijos añadidos para formar palabras son factores externos de dificultad a la hora de recuperar una palabra. La gramática también añade dificultad a la recuperación de palabras. Ej. la "s" de las terceras personas del singular de los verbos *makes*, *burns*, *tires* y *continues* o la terminación "–ing" en los verbos después de preposiciones; ej. g*enerating* en "new ways of generating" o *increasing* en "population goes on increasing".

Como ya explicamos en los adjetivos, las palabras cortas son más difíciles de recuperar, ya que la información proporcionada es muy escasa y puede dar lugar a la formación de varias palabras. Por ejemplo, las palabras *buy* (b__), *warn* (wa___), y *fear* (fe___), no están pautadas y son muy cortas por lo que aunque son palabras muy frecuentes en su vocabulario no han sido capaces de recuperarlas debido a la escasez de información. Los alumnos han dado otras palabras en lugar de *burns*, tales como *burnt*, *burst*, *build*, *built*, etc. o *fell*, *felt*, *feel*, etc. en lugar de *fear*, o *waste*, *watch*, *wake*, *warm*, etc. en lugar de *warn*, o *be* en lugar de *buy*.

El problema de *tires* es que es una palabra muy poco frecuente tanto en el BNC (34) como en el vocabulario de los alumnos. Ellos utilizan el término *tired* mucho más frecuentemente en expresiones como “to feel tired” o “ to be tired”, pero no el verbo *to tire*. Por ello han recuperado palabras más conocidas para ellos como son *tired*, *times*, *timer*, etc.

Cuando comprobamos la frecuencia de las palabras en el BNC hay que tener en cuenta la frecuencia de la clase de palabra que estemos analizando y no solamente la frecuencia general de esa palabra. Por ejemplo, la palabra *benefit* tiene una frecuencia de 10.767 en el BNC, pero solamente una frecuencia de 2.343 cuando actúa como verbo. Así pues, *benefit* es mucho más común como sustantivo que como verbo. Esto también hay que tenerlo en cuenta a la hora de evaluar la dificultad de recuperar una palabra y ésta puede ser una de las razones por la que solamente 31 de los 74 alumnos han podido recuperarla.

7.5. Adverbios

ADVERBIOS ≥ 60

Tabla 7.14

C-TEST A					
Adverbios ≥ 60	BNC	Correct.	Incor.	% Cor.	% Incor.
Actually	25429	64	13	83,1	16,9

Tabla 7.15

C-TEST B					
Adverbios ≥ 60	BNC	Correct.	Incor.	% Cor.	% Incor.
Probably	26484	65	8	89,2	10,8
Successfully	3337	62	12	83,8	16,2

Todos los adverbios que se han recuperado en los dos tests siguen la regla más común de formación de adverbios en inglés a partir de un adjetivo que es añadir el sufijo “ly” al adjetivo. En este caso también el 75% de los adverbios corresponden a textos pautados lo que les ha ayudado a restablecer las palabras mutiladas y a no cometer faltas de ortografía.

Observamos que el adverbio *probably* tiene una recuperación muy alta a pesar de corresponder a un texto no pautado. La explicación otra vez puede deberse a su gran similitud con la palabra en español y a la concurrencia de las tres palabras “it is probably” cuya frecuencia en inglés es bastante alta (478)

teniendo en cuenta que se trata de que estas tres palabras se usen conjuntamente.

Las palabras que los candidatos tenían que restaurar eran: “actu - - - -“, “succes - - - - - -“ y “prob_____”. La recuperación de estos adverbios puede haberse visto favorecida por el hecho de ser palabras largas con lo cual el número de letras que no se han mutilado, y por lo tanto las pistas que se han proporcionado a los alumnos, han sido también mayores.

ADVERBIOS ≤ 40

Tabla 7.16

C-TEST A					
Adverbios ≤ 40	BNC	Correct.	Incor.	% Cor.	% Incor.
Later	30735	32	45	41,6	58,4

De los tres adverbios que se encuentran en el C-test A: *actually, well* y *later*, el primero ha sido recuperado por un número elevado de alumnos, *well* a pesar de ser una palabra corta y no pautada la han reconstruido más de la mitad del grupo (46) debido a que es una palabra muy frecuente entre el léxico que ellos conocen. Por último, *later* ha sido recuperada por un número menor de cuarenta (32 alumnos exactamente). La palabra *later* se encuentra en la frase: “Regular exercise temporally ti - - - the body but la - - - on ...”. La mayoría de los alumnos, al darse cuenta de la preposición “on” después de la palabra mutilada, han pensado que la palabra que tenían que reconstruir se trataba de un verbo que expresaba contraste con otro verbo “ti - - - “ (tires*)*. Como puede observarse el verbo *tires* solo lo han reconstruido 14 personas (tabla 7.12) con lo que a la inmensa mayoría de los alumnos les faltaba información para poder reconstruir la palabra mutilada.

Todos los adverbios del C-test B terminan en “ly”, que como ya hemos visto es el sufijo que los alumnos generalmente asocian con un adverbio y que por lo tanto no han tenido mayores problemas para reconstruirlos. A ello puede deberse el que no exista ningún adverbio en el C-test B cuya recuperación haya sido igual o inferior a 40.

7.6. Conclusiones

Resumiendo todo lo anteriormente expuesto al analizar los resultados, podemos decir que el que una palabra se recupere más o menos fácilmente depende de numerosos factores tales como:

1. Si se considera correcta solamente la palabra exacta.

2. La frecuencia de la palabra en el uso del idioma tanto oral como escrito.

3. La frecuencia de la palabra en el vocabulario tanto activo como pasivo del alumno, que no tiene por qué coincidir con la frecuencia global de esa palabra en el idioma, ya que los estudiantes no aprenden necesariamente el vocabulario en el orden de frecuencia de las palabras en inglés.

4. La semejanza de esa palabra con otra palabra de igual significado en el primer idioma del alumno, en este caso el español.

5. La longitud de la palabra. Cuanto más corta sea la palabra más difícil será recuperarla, ya que se dispone de menor información. Ej. “fo_” (food), “n__” (new), “b__” (buy), “lo__” (look), “re___” (rest) o “we__” (well).

6. Si las palabras son pautadas o no pautadas. Las palabras pautadas se recuperan en un porcentaje más alto, ya que se dispone de mayor información.

7. Las flexiones y los sufijos añadidos para formar palabras que también son factores externos de dificultad a la hora de recuperar una palabra.

8. La dificultad del texto.

9. El nivel de competencia del alumno

Si analizamos el C-TEST A con más detenimiento, cuyos cuatro textos fueron ordenados de más fácil a más difícil, vemos que en el primero, PHYSICAL EXERCISE, de las 13 palabras de léxico 10 han sido recuperadas por un número de alumnos ≥ 60 (el 76,9%) y solamente dos palabras: *tires* y *later* han sido recuperadas por un número de alumnos ≤ 40 de los 77 que hicieron este modelo de test. Sin embargo, de las 12 palabras de léxico del texto VITAMINS ARE VITAL, que era el texto más difícil y estaba sin pautar, 8 de ellas (el 66,7%) sólo han sido recuperadas por un número de alumnos ≤ 40 y ninguna palabra ha sido recuperada por un número de alumnos ≥ 60.

Los resultados anteriores pueden indicar que cuando el texto es más difícil y además tienen que recuperar las palabras sin pautar, los alumnos pierden fácilmente el sentido del texto lo cual hace prácticamente imposible poder recuperar el léxico al carecer de la información textual necesaria. Como afirma Sigott (2004), cuando el nivel del alumno no es muy alto se requiere que el término esté completamente contextualizado para poder reconstruirlo, ya que en condiciones de contexto reducido las dificultades que encuentran son mayores.

Tenemos que considerar también que este modelo de prueba era totalmente nuevo para ellos y que el factor cansancio puede haber influido en que los peores resultados de recuperación de las palabras se hayan obtenido en el último texto. Así pues, los datos obtenidos reflejan que para un mismo

texto, los términos conocidos, sin flexiones, semejantes a términos del primer idioma del alumno, no excesivamente cortos y que estén pautados serán más fácilmente recuperados.

Estos resultados concuerdan con los hallados por Kontra y Kormos (2006) que al investigar las estrategias utilizadas por los participantes en la reconstrucción de las palabras mutiladas de un C-test encontraron que la mayoría de los alumnos, en principio, no utilizaron ninguna técnica y rellenaban los términos directamente. Cuando ya empezaron a utilizar alguna estrategia, la más común era la de contar el número de letras, buscar en su léxico mental y activar las palabras que podían encajar en el texto. Para ello utilizaban frecuentemente sus conocimientos sintácticos. También utilizaron las estrategias textuales y de traducción aunque de forma menos representativa. Las estrategias a las que prestaron menor atención fueron el uso de sus conocimientos del tema o de las pistas morfológicas que se les proporcionaba para reconstruir los términos. Tanto Anderson (1991) como Vann y Abraham (1990) sostienen que el éxito para reconstruir el C-test se basa en conocer no solamente qué estrategia hay que utilizar sino en saber también como utilizarla con éxito y como coordinar su uso con otras estrategias.

Capítulo 8

ESTUDIO EMPÍRICO DEL C-TEST

8.1. Observaciones generales

Los resultados que se muestran en la Tabla 8.1 son los datos descriptivos de los cuatro textos que forman el C-test, siendo *Ph*. (Physical exercise); *Re*. (Relax and live); En. (Alternative sources of energy); y *Vi.* (Vitamins are vital).

Un análisis de la media, la moda y la mediana de los cuatro textos muestra que la distribución de los valores es prácticamente normal. Los datos estadísticos de asimetría señalan, sin embargo, un sesgo ligeramente negativo en su distribución, especialmente en los primeros tres textos, lo que significa que el C-test no resultó una prueba difícil para la mayoría de los alumnos que sacaron buenas puntuaciones.

Tabla 8.1. Datos descriptivos [2]

		TtPh	TtRe	TtEn	TtVi
N	Válidos	151	151	151	151
	Perdidos	0	0	0	0
Media		18,4172	17,4238	17,5166	15,1722
Mediana		19,0000	18,0000	18,0000	15,0000
Moda		22,00	22,00	18,00	18,00
Desv. típ.		4,33952	4,75035	4,32181	4,42984
Asimetría		-,759	-,850	-,568	-,246
Error típ. de asimetría		,197	,197	,197	,197
Curtosis		,220	,624	-,228	-,116
Error típ. de curtosis		,392	,392	,392	,392
Rango		21,00	23,00	18,00	22,00
Mínimo		4,00	2,00	6,00	2,00
Máximo		25,00	25,00	24,00	24,00

La media y la mediana nos dan el grado de dificultad de los textos. Se puede apreciar que los textos fueron fáciles para los alumnos, ya que las

[2] TtPh: total physical exercise; TtRe: total relax and live TtEn: total alternative sources of energy TtVi total vitamins are vital

medias son: 18,42; 17,42; 17,52; y 15,17 (cuando la puntuación máxima que se podía obtener era 25) muy por encima del 50% ó 60% recomendado por los creadores del test. El primer texto es el más fácil y el último el más difícil, siendo los dos de en medio de dificultad similar. Esto significa que cuando ordenamos los textos la predicción del grado de dificultad se hizo correctamente.

Los resultados que se muestran en las tablas 8.2, 8.3 y 8.4 son los datos estadísticos descriptivos de los dos modelos de C-test (el C-test A y el C-test B) y del C-test global incluyendo el número de sujetos (N), la media, la desviación típica, la puntuación mínima, la puntuación máxima, y el rango. Hay que tener en cuenta que el número total de sujetos fue 151 de los cuales 77 hicieron el C-test A y 74 el C-test B.

Tablas 8.2. Estadísticos descriptivos: C-test A [3]

	N	Rango	Mínimo	Máximo	Media	Desv. típ.
F.Ph	77	8,00	4,00	12,00	10,5714	1,66566
L.Ph	77	13,00	,00	13,00	9,8961	2,06209
F.Re	77	14,00	2,00	16,00	11,5714	3,38895
L.Re	77	9,00	,00	9,00	4,7273	2,18630
F.En	77	9,00	2,00	11,00	7,7662	2,35023
L.En	77	12,00	2,00	14,00	8,8831	2,57505
F.Vi	77	11,00	2,00	13,00	9,7532	2,50345
L.Vi	77	11,00	,00	11,00	4,3117	2,37453
N válido (según lista)	77					

[3] F.Ph: función physical exercise; L..Ph: léxico physical exercise ; F. Re: función relax and live; L. Re: léxico relax and live; F. En: función alternative sources of energy; L. En: léxico alternative sources of energy; F. Vi: función vitamins are vital; L.Vi: léxico vitamins are vital.

Tablas 8.3. Estadísticos descriptivos: C-test B

	N	Rango	Mínimo	Máximo	Media	Desv. típ.
F.Ph	74	11,00	2,00	13,00	9,5270	2,39419
L.Ph	74	10,00	2,00	12,00	6,7568	2,50922
F.Re	74	15,00	1,00	16,00	12,8784	2,60112
L.Re	74	9,00	,00	9,00	5,7162	1,79436
F.En	74	9,00	4,00	13,00	10,5541	1,98723
L.En	74	10,00	2,00	12,00	7,8649	2,37755
F.Vi	74	10,00	2,00	12,00	7,6757	2,92419
L.Vi	74	9,00	4,00	13,00	8,6486	2,00979
N válido (según lista)	74					

Tablas 8.4. Estadísticos descriptivos: C-test global

	N	Rango	Mínimo	Máximo	Media	Desv. típ.
F.Ph	151	11,00	2,00	13,00	10,0596	2,11418
L.Ph	151	13,00	,00	13,00	8,3576	2,77451
F.Re	151	15,00	1,00	16,00	12,2119	3,08892
L.Re	151	9,00	,00	9,00	5,2119	2,05786
F.En	151	11,00	2,00	13,00	9,1325	2,58373
L.En	151	12,00	2,00	14,00	8,3841	2,52418
F.Vi	151	11,00	2,00	13,00	8,7351	2,90219
L.Vi	151	13,00	,00	13,00	6,4371	3,09101
N válido (según lista)	151					

8.2. Términos léxicos y funcionales

Para poder cuantificar los datos se ha dividido cada texto en dos subtests: subtest léxico y subtest funcional. Ello quiere decir que hemos creado 8 subtests (4 subtests léxicos y otros 4 funcionales). En lo que llamamos “subtets léxicos” analizamos exclusivamente los términos léxicos que deben ser recuperados en cada uno de los 4 textos y en los subtests funcionales hacemos lo mismo con los términos funcionales.

Analizando el C-test A observamos que existen diferencias significativas en los resultados de los alumnos especialmente en los términos léxicos. Puede verse que en tres de los cuatro subtests léxicos, el número mínimo de términos léxicos recuperados es cero mientras que en los subtests funcionales, aunque el número mínimo de recuperaciones pueda ser bajo, nunca es cero. Esto indica que, en general, los alumnos tienen más dificultades en recuperar los términos léxicos que los funcionales y que para alguno de ellos fue imposible restaurar ni un solo término léxico de alguno de los cuatro textos.

Si analizamos las medias de las palabras que han sido recuperadas en cada uno de los subtests nos puede llamar la atención que la media de *F. En* (7,7662) es menor que la media de *L. En* (8,8831), lo cual parece contradecir lo expuesto anteriormente. Sin embargo, debido a que el número de términos léxicos y funcionales no es el mismo en todos los subtests, calculamos el porcentaje de palabras léxicas y funcionales recuperadas en ese texto del modelo A y comprobamos que se han recuperado el 70,6% de los términos funcionales y el 63,5% de los léxicos. Esto nos demuestra que en este texto también se cumple la regla de que las palabras funcionales se recuperan más fácilmente que las léxicas, aunque a primera vista no pareciera ser así.

La diferencia de medias que existe entre los términos funcionales y los léxicos del texto “Relax and Live” del modelo A (F. Re. = 11,5714 y L. Re. = 4,7273) es manifiesta. Ello es debido también a que el número de palabras funcionales que se han mutilado en este texto es de 16 mientras que el número de palabras léxicas es sólo de 9. Calculando el porcentaje de términos funcionales recuperados vemos que es del 72,3% frente al 52,5% de los léxicos.

Por otra parte, podemos observar que hay alumnos que han recuperado todos los términos, lo que quiere decir que el rango de respuestas ha sido amplio. La dispersión puede verse por el rango, el cual es el máximo posible en varios de los subtests tanto de léxico como de función.

En el C-test B vemos que solamente en el subtest léxico de “Relax and Live” ha habido algún alumno que no ha podido recuperar ninguno de los términos léxicos mientras que en el C-test A esto ocurría en tres de los cuatro subtests. En este mismo texto de “Relax and Live” vemos que el número de palabras léxicas y funcionales que los alumnos tienen que recuperar en el modelo A es 9 y 16 respectivamente, igual que las que tienen que recuperar en el mismo texto del modelo B. Sin embargo, observamos que las respuestas han variado ligeramente. Mientras que en el modelo A se han recuperado el 72,3% de los términos funcionales y el 52,5% de los léxicos, en el modelo B el porcentaje de términos recuperados en cada subtest ha aumentado. Así vemos que se han recuperado el 80,5% de los términos funcionales y el 63,5% de los léxicos. Esto puede verse más claramente en el siguiente cuadro comparativo: Tabla 8.5

Tabla 8.5. Cuadro comparativo de recuperación de palabras en dos de los 8 subtests creados para cada modelo de C-test.

Subtests	N[4]	% Recuperación	
		C-test A	C-test B
F. Re.[5]	16	72,3	80,5
L. Re[6].	9	52,5	63,5

La homogeneidad de los dos grupos de alumnos se demostró con la aplicación del test del Tutor en el que se vio que no había diferencias significativas entre los dos grupos. Por otra parte, variables como el contenido o el tema del pasaje, la densidad del texto o la dificultad lingüística no pueden ser consideradas, ya que el texto base con el que se elaboraron los subtests “Relax and Live” en los dos modelos C-tests A y C-test B es el mismo. Además, el número de términos léxicos y funcionales que tenían que recuperar los alumnos es idéntico en los dos modelos de test. Por lo tanto podemos afirmar

[4] Número de términos mutilados en los subtests A y B del texto Relax and Live.
[5] Subtests de Función Relax and Live.
[6] Subtests de léxico Relax and Live.

que la diferencia que existe en el porcentaje de recuperación de palabras en ambos modelos tiene que deberse exclusivamente a las palabras concretas que tienen que ser recuperadas en cada caso. Es decir, que el hecho de que las palabras que haya que recuperar en ambos modelos de test sean distintas afecta a los resultados.

En la tabla 8.6. se pueden apreciar los porcentajes de recuperación de los términos léxicos de los diferentes subtests que forman los dos modelos A y B de C-test.

Tabla 8.6. Cuadro comparativo de recuperación de los términos léxicos en los dos modelos de C-test.

Subtests[7]	% Recuperación	
	C-test A	C-test B
L. Ph.	76,1	56,3
L. Re.	52,5	63,5
L. En.	63,4	65,5
L. Vi.	35,9	68,2

La diferencia entre los términos léxicos recuperados en alguno de los subtests del C-test A y C-test B es considerable y sorprendente. Así vemos que en el subtest *L. Ph.* existe una diferencia de prácticamente un 20% y en el subtest L. Vi. la diferencia es de un 32,3%. Basándonos en todos los datos anteriores empezamos a intuir que el punto donde se comience a mutilar el texto va a afectar a las puntuaciones obtenidas, por lo menos a nivel de texto o de subtests.

En cuanto a las medias de los subtests del C-test Global, éstas siguen la misma pauta que las de los tests individuales, C-test A y C-test B, corroborando

[7] Subtests de Léxico: L.Ph: léxico physical exercise; L. Re: léxico relax and live; L. En: léxico alternative sources of energy; L.Vi: léxico vitamins are vital.

así los resultados anteriores que establecían que los términos funcionales se recuperan más fácilmente que los léxicos. Tablas 8.2, 8.3, 8.4,y 8.7.

Tabla 8.7. Porcentaje de recuperación de los términos léxicos y funcionales en el C-test Global.

subtests		Media	% Recuperación
1	F. Ph.	10,0596	80,48
	L. Ph.	8,3576	66,86
2	F. Re.	12,2119	76,32
	L. Re.	5,2119	57,91
3	F. En.	9,1325	76,10
	L. En.	8,3841	64,49
4	F. Vi.	8,7351	69,88
	L. Vi.	6,4371	51,50

En esta tabla se observa claramente que en todos y cada uno de los textos que forman el C-test global el porcentaje de recuperación de los términos funcionales (representados por los subtests funcionales) es mayor que el de los términos léxicos (representados por los subtests léxicos).

En los siguientes apartados vamos a determinar si esas diferencias son estadísticamente significativas o no. En los resultados que se muestran en las tablas 8.7 y 8.8 puede observarse que para cada texto la media de los subtests funcionales es siempre mayor que la media de los subtests léxicos y que lo mismo ocurre con los resultados globales del C-test funcional (C-Test Function) y los del C-test léxico (C-Test Lexis). Por lo tanto, podemos afirmar que los subtests funcionales o de estructuras son más fáciles que los subtests léxicos o de contenido y también que el C-test funcional global es más fácil que el C-test léxico global.

Tabla 8.8. Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	F.Ph	10,0596	151	2,11418	,17205
	L.Ph	8,3576	151	2,77451	,22579
Par 2	F.Re	12,2119	151	3,08892	,25137
	L.Re	5,2119	151	2,05786	,16747
Par 3	F.En	9,1325	151	2,58373	,21026
	L.En	8,3841	151	2,52418	,20541
Par 4	F.Vi	8,7351	151	2,90219	,23618
	L.Vi	6,4371	151	3,09101	,25154
Par 5	C-Test. Lexis	28,3907	151	7,06066	,57459
	C-Test. Function	40,1391	151	8,27771	,67363

Como las medias de los pares no son independientes, ya que los mismos estudiantes han hecho los dos tests, se ha usado el t-test para medias correlacionadas para comparar las medias entre los resultados de los términos léxicos y funcionales que han sido restaurados.

Tabla 8.9. Prueba de muestras relacionadas

		t	gl	Sig. (bilateral)
Par 1	F.Ph - L.Ph	8,915	150	,000
Par 2	F.Re - L.Re	38,520	150	,000
Par 3	F.En - L.En	3,377	150	,001
Par 4	F.Vi - L.Vi	6,988	150	,000
Par 5	C-Test. Lexis - C-Test. Function	-28,131	150	,000

Se ha efectuado la comparación bilateral, usando el t-test, con la hipótesis nula de que las medias de las palabras funcionales y de contenido serían iguales con $p < 0,05$.

Para $gl = 150$ y $p < 0,05$ el valor crítico de $t = 1,96$. Como puede observarse en la tabla 8.9 todos los valores de $t_{observados}$ son superiores a 1,96.

Esto significa que existen diferencias muy significativas ($p < 0,01$) entre las medias de los términos funcionales y las de los léxicos y que por lo tanto podemos afirmar que las palabras funcionales son mucho más fáciles de recuperar que las de contenido. Se observan también diferencias significativas no sólo en todos y cada uno de los super-ítems sino también en el C-test global. De hecho, los datos de esta investigación nos confirman que se han recuperado el 75,73% de las palabras funcionales y el 60,40% de las de léxico.

Estos resultados están de acuerdo con la teoría que dice que la probabilidad de restauración de las palabras pertenecientes a una clase es, en general, inversamente proporcional al tamaño de la misma. Por lo tanto, las palabras funcionales son altamente predecibles, puesto que pertenecen a clases pequeñas y cerradas (Klein- Braley, 1985).

Por otra parte, aunque el número absoluto de palabras funcionales es pequeño, sin embargo, son muy frecuentes en cuanto a su uso. Esto puede apreciarse claramente en los histogramas (figuras 8.1 y 8.2). El histograma que muestra la recuperación de los términos funcionales tiene un sesgo negativo lo cual indica que a la mayoría de los alumnos que han realizado el test les ha resultado relativamente fácil recuperar los términos de función. Sin embargo, el histograma que muestra la recuperación de los términos léxicos ofrece una distribución más aproximada a la curva normal.

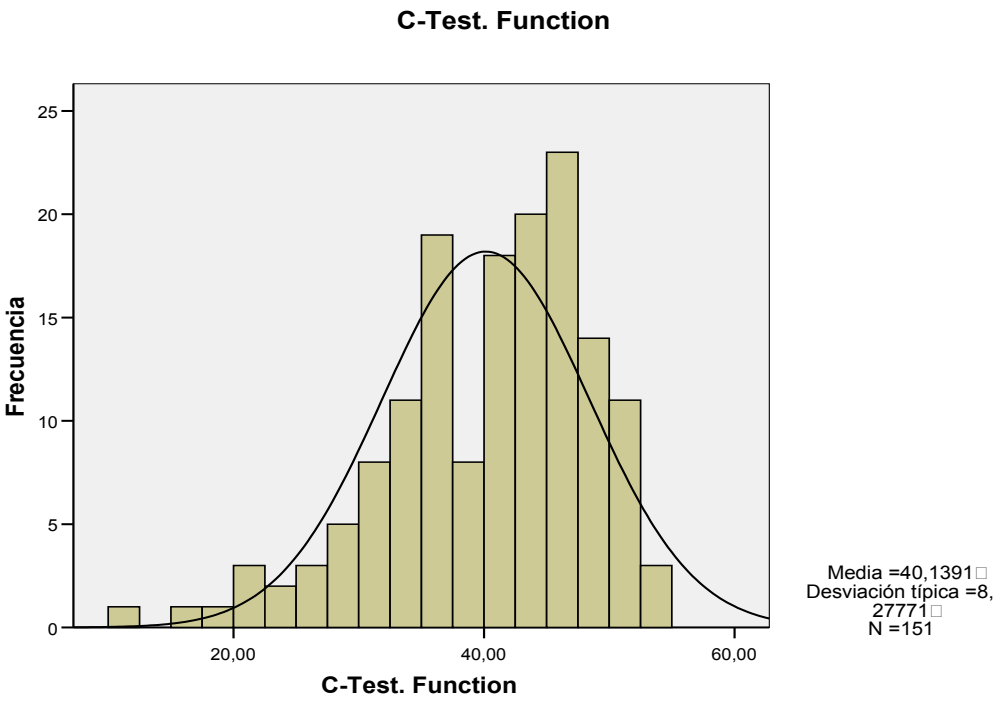

Figura 8.1. Histograma de la recuperación de los términos de función

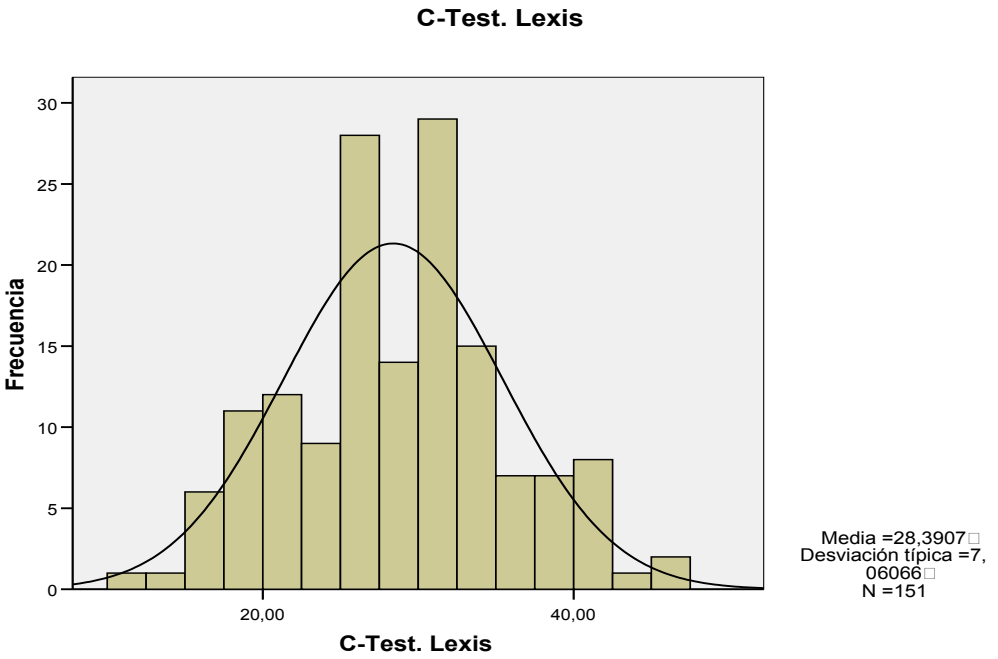

Figura 8.2. Histograma de la recuperación de los términos léxicos

8.3. Influencia del proceso de mutilación y homogeneidad de los grupos

También queríamos estudiar si el punto en que se empezaba a mutilar las palabras afectaba a los resultados de los alumnos. Para ello hemos comparado las medias de los diferentes pares de subtests de los dos modelos de C-test. (Tablas 8.10, 8.11, 8.12 y 8,13).

El examen de los resultados (tablas 8.10 y 8.11) muestra que las diferencias de medias de los subtests A y B son significativas. Como ya vimos en la tabla 8.6, las mayores diferencias se muestran en los subtests léxicos especialmente en los de los textos *Physical Exercise* (L. Ph.) y *Vitamins are Vital* (L. Vi.). Sin embargo, la media del test del Tutor parece ser muy similar tanto para el grupo A como para el grupo B.

Tabla 8.10. Estadísticos de grupo

	Tipo	N	Media	Desviación típ.	Error típ. de la media
F.Ph	A	77	10,5714	1,66566	,18982
	B	74	9,5270	2,39419	,27832
L.Ph	A	77	9,8961	2,06209	,23500
	B	74	6,7568	2,50922	,29169
F.Re	A	77	11,5714	3,38895	,38621
	B	74	12,8784	2,60112	,30237
L.Re	A	77	4,7273	2,18630	,24915
	B	74	5,7162	1,79436	,20859
F.En	A	77	7,7662	2,35023	,26783
	B	74	10,5541	1,98723	,23101
L.En	A	77	8,8831	2,57505	,29345
	B	74	7,8649	2,37755	,27638
F.Vi	A	77	9,7532	2,50345	,28529
	B	74	7,6757	2,92419	,33993
L.Vi	A	77	4,3117	2,37453	,27060
	B	74	8,6486	2,00979	,23363
Tutor	A	77	8,351	3,0077	,3428
	B	74	8,797	3,1448	,3656

De nuevo se utilizó el t-test para comparar las medias de los subtests y las medias del test del Tutor en ambos grupos A y B. Puede observarse que para los tests del Tutor las medias son prácticamente las mismas (tabla 8.10) y que las $t_{observadas} < t_{crit.}$, con $t_{crit} = 1{,}96$ y $t_{observadas} = 0{,}892$ con un nivel de significación de 0,374 (tabla 8.11). Ello quiere decir que la hipótesis nula no puede ser descartada y tenemos que aceptar que no existe diferencia significativa entre las medias de los resultados del test del Tutor de los dos grupos. Estos resultados confirman la homogeneidad de la muestra y por lo tanto podemos afirmar que las diferencias significativas que se muestran en las tablas 8.10, y 8.11 se deben exclusivamente a las diferencias existentes entre los dos modelos de C-test.

Tabla 8.11. Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
									Superior	Inferior
F.Ph	Se han asumido varianzas igua Les	14,503	,000	3,122	149	,002	1,04440	,33455	,38332	1,70548
	No se han asumido varianzas iguales			3,100	129,743	,002	1,04440	,33689	,37790	1,71090
L.Ph	Se han asumido varianzas iguales	4,532	,035	8,414	149	,000	3,13935	,37313	2,40204	3,87665
	No se han asumido varianzas iguales			8,381	141,327	,000	3,13935	,37458	2,39885	3,87984
F.Re	Se han asumido varianzas iguales	9,721	,002	-2,651	149	,009	-1,30695	,49304	-2,28120	-,33270
	No se han asumido varianzas iguales			-2,665	142,130	,009	-1,30695	,49050	-2,27656	-,33734
L.Re	Se han asumido varianzas iguales	3,637	,058	-3,032	149	,003	-,98894	,32621	-1,63354	-,34435
	No se han asumido varianzas iguales			-3,043	145,472	,003	-,98894	,32494	-1,63116	-,34673
F.En	Se han asumido varianzas iguales	4,876	,029	-7,856	149	,000	-2,78782	,35487	-3,48906	-2,08658
	No se han asumido varianzas iguales			-7,882	146,645	,000	-2,78782	,35370	-3,48682	-2,08882
L.En	Se han asumido varianzas iguales	,080	,777	2,522	149	,013	1,01825	,40376	,22042	1,81609
	No se han asumido varianzas iguales			2,526	148,765	,013	1,01825	,40312	,22168	1,81483
F.Vi	Se han asumido varianzas iguales	4,822	,030	4,696	149	,000	2,07757	,44242	1,20334	2,95180
	No se han asumido varianzas iguales			4,681	143,616	,000	2,07757	,44379	1,20038	2,95477
L.Vi	Se han asumido varianzas iguales	1,120	,292	-12,091	149	,000	-4,33696	,35869	-5,04574	-3,62819
	No se han asumido varianzas iguales			-12,131	146,681	,000	-4,33696	,35751	-5,04349	-3,63043
Tutor	Se han asumido varianzas iguales	,146	,703	-,892	149	,374	-,4466	,5007	-1,4360	,5427
	No se han asumido varianzas iguales			-,891	147,944	,374	-,4466	,5011	-1,4369	,5437

Las Tablas 8.12 y 8.13 muestran los análisis estadísticos de los cuatro super-ítems (los 4 textos) tanto del C-test A como del C-test B.

Tabla 8.12. Estadísticos de grupo

	Tipo	N	Media	Desviación típ.	Error típ. de la media
TtPh	A	77	20,4675	3,05900	,34861
	B	74	16,2838	4,46146	,51863
TtRe	A	77	16,2987	5,21401	,59419
	B	74	18,5946	3,91666	,45530
TtEn	A	77	16,6494	4,60482	,52477
	B	74	18,4189	3,83212	,44547
TtVi	A	77	14,0649	4,19692	,47828
	B	74	16,3243	4,39834	,51130
C.Test	A	77	67,4805	14,85687	1,69310
	B	74	69,6216	14,14863	1,64474

Tabla 8.13. Prueba de muestras independientes

		Prueba T para la igualdad de medias						
		t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
							Inferior	Superior
Ttph	Se han asumido varianzas iguales	6,743	149	,000	4,18375	,62042	2,95780	5,40970
	No se han asumido varianzas iguales	6,695	128,642	,000	4,18375	,62491	2,94732	5,42017
TtRe	Se han asumido varianzas iguales	-3,050	149	,003	-2,29589	,75276	-3,78335	-,80844
	No se han asumido varianzas iguales	-3,067	140,883	,003	-2,29589	,74857	-3,77579	-,81600
TtEn	Se han asumido varianzas iguales	-2,561	149	,011	-1,76957	,69086	-3,13471	-,40442
	No se han asumido varianzas iguales	-2,571	146,044	,011	-1,76957	,68835	-3,12999	-,40915
TtVi	Se han asumido varianzas iguales	-3,230	149	,002	-2,25939	,69947	-3,64156	-,87722
	No se han asumido varianzas iguales	-3,227	147,886	,002	-2,25939	,70013	-3,64294	-,87584
C.Test	Se han asumido varianzas iguales	-,906	149	,366	-2,14110	2,36276	-6,80995	2,52775
	No se han asumido varianzas iguales	-,907	148,988	,366	-2,14110	2,36046	-6,80540	2,52320

Se descarta la hipótesis nula de que el punto donde se comienza a mutilar las palabras no tiene ningún efecto sobre los resultados de los alumnos a nivel de super-ítems, ya que se ha observado que existen diferencias significativas entre las medias de todos y cada uno de los super-ítems (tablas 8.12 y 8.13). Es decir, que cambiando el punto donde se comienza a mutilar las palabras puede dar lugar a diferentes super-ítems.

Sin embargo, como muestran los análisis estadísticos, la hipótesis nula que considera que el procedimiento de supresión de palabras en los dos modelos de C-test no tiene ningún efecto, no puede ser desechada, ya que se observa un valor de t= 0,906 con un nivel de significación de 0,366 (tablas 8.13 y 8.15). Esto quiere decir que no existen diferencias significativas entre los dos modelos de C-test que hemos creado a partir de cuatro textos diferentes con 25 palabras mutiladas en cada uno de ellos. Si no existen diferencias significativas entre los dos modelos de C-test podemos decir que el C-test A es equivalente al C-test B.

Las tablas 8.14 y 8.15 resumen el hallazgo más significativo de este trabajo de investigación que es que los dos C-tests que se han construido a partir de unos mismos textos, siguiendo las indicaciones de sus creadores Klein-Braley y Raatz, son tests equivalentes o paralelos.

Tabla 8.14. Estadísticos de grupo

	Tipo	N	Media	Desviación típ.	Error típ. de la media
Tutor	A	77	8,351	3,0077	,3428
	B	74	8,797	3,1448	,3656
C-Test	A	77	67,4805	14,85687	1,69310
	B	74	69,6216	14,14863	1,64474

Tabla 8.15. Prueba de muestras independientes

		Prueba T para la igualdad de medias						
		t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
							Superior	Inferior
Tutor	Se han asumido varianzas iguales	-,892	149	,374	-,4466	,5007	-1,4360	,5427
	No se han asumido varianzas iguales	-,891	147,944	,374	-,4466	,5011	-1,4369	,5437
C-Test	Se han asumido varianzas iguales	-,906	149	,366	-2,14110	2,36276	-6,80995	2,52775
	No se han asumido varianzas iguales	-,907	148,988	,366	-2,14110	2,36046	-6,80540	2,52320

Podemos hacer la anterior afirmación debido a que los dos modelos de C-test, el C-test A y el C-test B, los han realizado dos grupos de alumnos diferentes pertenecientes al mismo nivel de la EOI, y que como podemos ver en las tablas 8.14 y 8.15 son grupos homogéneos, ya que no existen diferencias significativas entre las medias de los resultados del test del tutor, el cual ha sido utilizado para comprobar esa homogeneidad. Estos resultados confirman la hipótesis de que los dos grupos son verdaderamente comparables en términos de competencia lingüística.

Resumiendo, todos los datos anteriores son de gran importancia, ya que podemos afirmar que por una parte, el C-test A es equivalente al C-test B (tablas 8.14 y 8.15) y por otra, hemos visto que existen diferencias significativas entre todos los super-ítems A y B de los dos C-tests (tablas 8.10, 8.11, 8.12 y 8.13). Como se ha demostrado que los dos grupos de alumnos que hicieron el examen componen una muestra homogénea (tablas 8.14 y 8.15), la única diferencia que existe entre el C-test A y el C-test B es el punto donde se empieza a mutilar los textos. Por lo tanto, podemos llegar a la conclusión de que el comienzo de mutilación de los textos crea subtests diferentes lo cual afecta a los resultados de los mismos pero no afecta a los de los dos modelos de C-test creados, que como ya se ha demostrado son equivalentes.

En esta investigación se ha trabajado con textos cortos, ya que se ha construido el C-test de acuerdo a las directrices de sus creadores. Sin embargo, los resultados anteriores nos llevan a reflexionar y a pensar que si esto ocurriera con cualquier tipo de texto, entonces sería totalmente desaconsejable construir un C-test con un solo texto, aunque fuera un texto extenso, ya que el C-test sería diferente dependiendo del punto donde se empezara a mutilarlo y por lo tanto los resultados también serían diferentes. Esto es lo que le ocurrió a Jafarpur (1995) que elaboró varios C-tests basándose en un solo texto largo del cual mutiló 125 palabras (Klein-Braley y Raatz limitan a 100 el número de supresiones de cada C-tests que se elabore con varios textos). Estas palabras eran distintas en cada modelo de C-test, ya que comenzó a mutilar las palabras en distintos puntos del texto y además variaba las ratios entre los términos mutilados. De ahí que llegara a la conclusión de que el C-test no era un método válido, ya que con un solo texto se podían crear varios tests diferentes. Ello quería decir que el C-test no era un test superior al *cloze* puesto que no lograba resolver los problemas que planteaba éste.

Sin embargo, según muestran los datos estadísticos (tablas 8.13 y 8.15) no existen diferencias significativas entre el C-test A y el C-test B. Si estos dos tests son equivalentes y cada uno de los subtests que los componen no lo son entre si, ello quiere decir que las diferencias existentes entre los subtests se neutralizan ente si y quedan anuladas al formar el C-test.

De nuevo este hallazgo tiene su importancia por varias razones:

1. Nos demuestra la importancia que tiene el utilizar varios textos a la hora de elaborar un C-test de acuerdo con sus diseñadores Klein-Braley y Raatz.

2. Corrobora la importancia que tiene el mutilar una de cada dos palabras y no utilizar otra ratio distinta, ya que de este modo podremos obtener dos C-tests diferentes con cada batería de textos y ahora sabemos que esos dos modelos de tests son

equivalentes por lo que no tendremos que preocuparnos del punto donde empecemos a mutilar las palabras.

3. Gracias a este hallazgo podemos afirmar que el C-test es superior al *cloze* por lo menos en su construcción y en su corrección, ya que estos dos procesos son mucho más fáciles en el C-test.

8.4. Comparación entre el *cloze* y el C-test

Tabla 8.16. Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	RCL	7,3344	151	1,68298	,13696
	C-test	6,1689	151	1,88572	,15346

La tabla 8.16 nos muestra las medias de los dos tests y vemos que la media del *cloze* es más alta que la del C-test. Esto puede ser debido a la falta total de experiencia con el procedimiento del C-test por parte de los alumnos, mientras que la técnica del *cloze* se viene incluyendo en la batería de exámenes de la EOI desde hace años con lo cual forma parte también de la práctica habitual en clase, por lo que no es extraño que les resulte más fácil.

Por otra parte, para resolver el *cloze* se proporciona a los alumnos un banco de respuestas además de unos términos extras como distractores. Ello quiere decir que el test era principalmente de conocimiento pasivo, por lo tanto, los examinandos no tienen que reproducir ningún término con lo que no es posible cometer errores gramaticales, morfológicos o de ortografía que les resten puntos.

Tabla 8.17. Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	RCL y C-test	151	,562	,000

En la tabla 8.17 observamos que la correlación entre el *cloze* y el C-test es bastante alta (0,562), lo cual es lógico, ya que los dos tests son técnicas de evaluación semejantes con lo cual el constructo que miden debe ser parecido.

Tabla 8.18. Prueba de muestras independientes

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
				Superior	Inferior			
Par 1 RCL - C.test	1,16556	1,68027	,13674	,89538	1,43575	8,524	150	,000

Para analizar las diferencias de las medias de los dos tests y determinar si son significativas o no, se considera que para gl = 150 y $p < 0,05$ la $t_{crit} = 1,96$. Como puede observarse en la tabla 8.18 el $t_{observado} = 8,524$ superior a 1,96. Esto significa que existen diferencias muy significativas ($p < 0,001$) entre la media de los resultados del *cloze* y la del C-test y que por lo tanto podemos afirmar que realmente el *cloze* ha sido más fácil que el C-test lo cual se aprecia claramente en los dos histogramas (figuras 8.3 y 8.4) de estos tests.

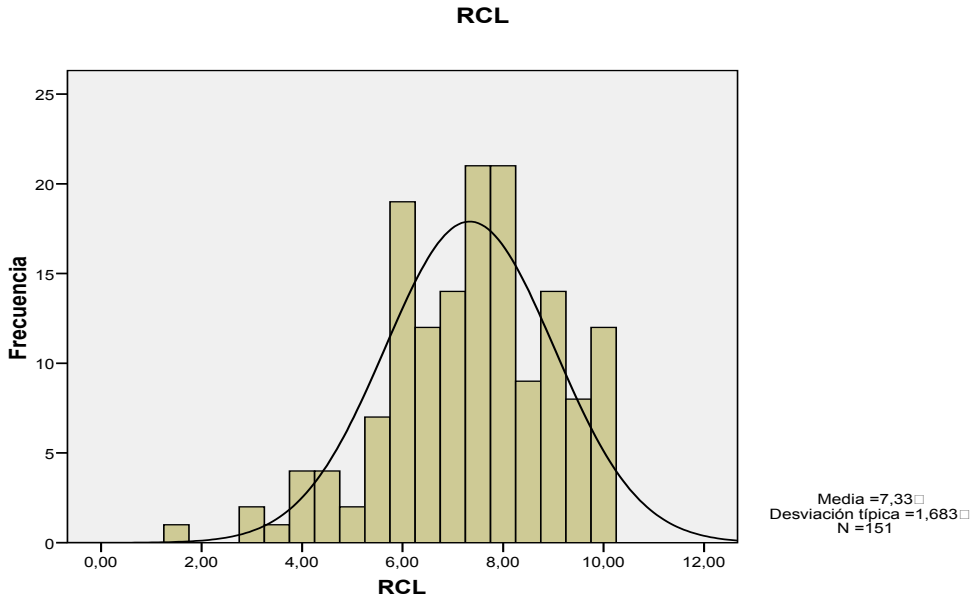

Figura 8.3. Histograma del *cloze*

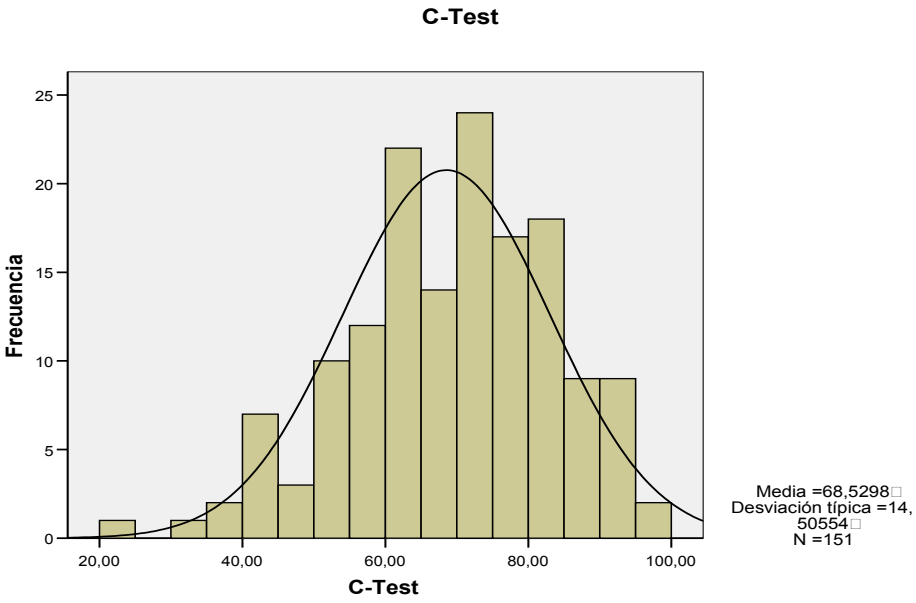

Figura 8.4. Histograma del C-test global

8.5. Subtests pautados y no pautados

Como se dijo en el apartado de materiales de este capítulo, tanto en el modelo C-test A como en el C-test B los huecos de los super-ítems "*Physical exercise*" y "*Alternative Sources of Energy*" se pautaron, es decir, se sustituyó cada letra suprimida por un guión. En cambio en los otros dos textos de los cuatro que forman el C-test: "*Relax and Live*" y "*Vitamins are Vital*", los huecos no se pautaron sino que las letras suprimidas se reemplazaron por un único guión. Se quiso analizar si esto tenía alguna influencia sobre los términos recuperados en cada super-ítem.

Al analizar los datos estadísticos de la tabla 8.19 observamos que existe una alta correlación (0,727) entre los subtests pautados y no pautados.

Tabla 8.19. Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	Pautados y NO Pautados	151	,727	,000

En la tabla 8.20 vemos que la media de los términos recuperados en los dos subtests que están pautados es superior a la de los dos subtests que no lo están. Lo mismo demuestra la tabla 8.21 donde vemos que la media global de términos recuperados es mayor entre los subtests pautados que entre los no pautados

Tabla 8.20. Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	TtPh	18,4172	151	4,33952	,35315
	TtEn	17,5166	151	4,32181	,35170
Par 2	TtRe	17,4238	151	4,75035	,38658
	TtVi	15,1722	151	4,42984	,36050

Tabla 8.21. Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	Pautados	35,9338	151	7,29627	,59376
	NO Pautados	32,5960	151	8,30917	,67619

Tabla 8.22. Prueba de muestras relacionadas

		Diferencias relacionadas					t	gl	Sig. (bilateral)
		Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
					Superior	Inferior			
Par 1	Paut. – No Pautados	3,33775	5,84339	,47553	2,39815	4,27735	7,019	150	,000

La tabla 8.22 nos muestra que esas diferencias de medias son significativas, ya que el $t_{observado} > t_{crit;}$ t = 7,019 con un gl =150 y un grado de significación de 0,000.

Así pues, podemos afirmar, como ya se ha demostrado en otros estudios (Babaii y Moghaddam, 2006; y Süssmilch,1984), que es más fácil recuperar una palabra cuando está pautada que cuando no lo está.

8.6. Fiabilidad de los tests de la EOI

A continuación vamos a analizar la fiabilidad y la correlación entre las tres destrezas de la EOI: comprensión lectora, comprensión auditiva y expresión escrita, (tablas 8.23, 8.24, y 8.25) así como la fiabilidad y la correlación entre las tareas que componen la batería de exámenes de la

comprensión lectora: *headings*, *test de elección múltiple,* y *cloze* (tablas 8.26, 8.27 y 8.28).

8.6.1. Fiabilidad del test global de la EOI

La tabla 8.23 nos muestra un coeficiente de fiabilidad del test de la EOI de 0,729, lo cual nos indica un alto grado de consistencia interna entre todos los tests que lo forman.

Tabla 8.23. Estadísticos de fiabilidad

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,729	,735	3

Tabla 8.24. Matriz de correlaciones entre elementos[8]

	Reading	Listening	Writing
Reading	1,000	,410	,551
Listening	,410	1,000	,482
Writing	,551	,482	1,000

Tabla 8.25. Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento
Reading	29,7881	39,887	,556	,615
Listening	33,0662	36,598	,490	,717
Writing	34,5166	43,651	,607	,578

[8] Reading: batería de tests de Comprensión Lectora; Listening: bateríade tests de Comprensión Auditiva; Writing: batería de tests de expresión escrita.

El test que mayor contribución aporta los resultados finales del test EOI es el "Writing" y el que menos contribuye es el "Listening" (tabla 8.25). Esto puede ser debido a que el "Writing" es una técnica integradora de conocimiento activo que coincide más ampliamente con la competencia global de una lengua, algo que no ocurre con el "Listening".

8.6.2. Fiabilidad del test de comprensión lectora de la EOI

Tabla 8.26. Estadísticos de fiabilidad: Reading

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,493	,494	3

Tabla 8.27. Matriz de correlaciones entre elementos[9]

	RH	RMC	RCL
RH	1,000	,057	,440
RMC	,057	1,000	,240
RCL	,440	,240	1,000

La tabla 8.26 nos muestra que el coeficiente de fiabilidad de los tres tests que componen el examen de "Reading" es moderado. En la tabla 8.27 vemos también que la correlación entre esos mismos tests no es muy alta, especialmente la correlación entre el test de elección múltiple y el de *headings*, lo cual nos indica que la consistencia interna entre los tests que componen la batería del examen de comprensión lectora debería revisarse. También indica

[9]Tests que forman la batería de tests de la Comprensión Lectora: RH: Test de casar elementos o "Headings"; RMC: Test de elección múltiple y RCL: cloze test.

que el solapamiento entre los constructos de estas tres pruebas es muy bajo por lo que cada prueba está midiendo un aspecto diferente de la destreza de comprensión lectora. Esto concuerda con lo dicho en el capítulo 3 (Weir, 2005a: 32) en el que se expresa que si la destreza de la comprensión lectora es divisible entonces no se puede esperar altos valores de consistencia interna entre los tests que evalúan esta destreza.

De acuerdo con Weir, los resultados obtenidos en nuestro estudio nos estarían indicando que efectivamente la destreza de la comprensión lectora es divisible y que el valor de correlación entre el test de *headings* y el de elección múltiple es tan bajo por estar midiendo distintos aspectos de la comprensión lectora.

Tabla 8.28. Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento
RH	13,0265	7,023	,316	,387
RMC	13,2053	8,908	,170	,610
RCL	11,5629	6,541	,473	,107

Como ya hemos visto anteriormente, el test de elección múltiple es el que peor correlaciona con los otros tests del “Reading” y según se observa en la tabla 8.28 el que menos contribuye a este examen. Por el contrario, el *cloze* es el test que más contribuye al examen de “Reading” y también el que tiene correlaciones más altas con el resto de los tests que componen el examen de comprensión lectora.

Este hallazgo nos ha sorprendido, ya que el *cloze* es uno de los tests que más oposición origina tanto entre los profesores como entre los alumnos, mientras que el test de elección múltiple es una de las técnicas que mayor aceptación tiene y que se utiliza con más frecuencia en los exámenes de las escuelas Oficiales de Idiomas. Esto nos lleva a pensar que no siempre la

fiabilidad y validez de un test coincide con la validez aparente y la aceptación del mismo por parte de los docentes y alumnos, lo cual coincide con lo que afirma Stevenson (1985).

8.7. Validez del C-test

8.7.1. Validez del contenido

Definimos la validez del contenido del C-test en función de si la distribución de las clases de palabras entre los términos que tienen que recuperar los alumnos representan una distribución normal de las clases de palabras en el texto original.

Para este estudio se elaboraron dos C-tests a partir de cuatro textos. En los dos primeros subtests del C-test A se empezó a mutilar las palabras a partir de la segunda palabra después del primer punto y en los dos primeros subtests del C-test B a partir de la tercera palabra después del primer punto. En los dos últimos subtests la mutilación se empezó en la tercera palabra después del primer punto en el C-test A y en la segunda palabra después del primer punto en el C-test B.

Una vez analizados los términos de léxico y de contenido que los alumnos tienen que recuperar en los dos modelos de C-test, comprobamos que no existen diferencias significativas entre ambos tests, ya que en el C-test A de los 100 términos que tienen que recuperar 52 son palabras funcionales y en el C-test B el número de términos funcionales que tienen que recuperar son 54, por lo que podemos afirmar que no existen desviaciones significativas.

Por otra parte, como ya se ha demostrado empíricamente, no existen diferencias significativas en la dificultad de los dos C-tests, ya que pueden considerarse tests paralelos, de ahí que midan el mismo constructo.

8.7.2. Validez relacionada con el criterio: validez concurrente

Las tablas 8.29 y 8.30 muestran la correlación entre los dos modelos del C-test y los diferentes tests de la EOI para establecer la validez de los C-tests.

Tabla 8.29. Correlaciones: C-test A

		C-test A	Reading	Listening	Writing	EOI
C-test A	Correlación de Pearson	1	,581(**)	,547(**)	,369(**)	,644(**)
	Sig. (bilateral)		,000	,000	,001	,000
	N	77	77	77	77	77
Reading	Correlación de Pearson	,581(**)	1	,377(**)	,520(**)	,795(**)
	Sig. (bilateral)	,000		,001	,000	,000
	N	77	77	77	77	77
Listening	Correlación de Pearson	,547(**)	,377(**)	1	,408(**)	,789(**)
	Sig. (bilateral)	,000	,001		,000	,000
	N	77	77	77	77	77
Writing	Correlación de Pearson	,369(**)	,520(**)	,408(**)	1	,779(**)
	Sig. (bilateral)	,001	,000	,000		,000
	N	77	77	77	77	77
EOI	Correlación de Pearson	,644(**)	,795(**)	,789(**)	,779(**)	1
	Sig. (bilateral)	,000	,000	,000	,000	
	N	77	77	77	77	77

** La correlación es significativa al nivel 0,01 (bilateral).

Tabla 8.30. Correlaciones: C-test B

		C-test B	Reading	Listening	Writing	EOI
C-test B	Correlación de Pearson	1	,579(**)	,491(**)	,521(**)	,642(**)
	Sig. (bilateral)		,000	,000	,000	,000
	N	74	74	74	74	74
Reading	Correlación de Pearson	,579(**)	1	,436(**)	,577(**)	,800(**)
	Sig. (bilateral)	,000		,000	,000	,000
	N	74	74	74	74	74
Listening	Correlación de Pearson	,491(**)	,436(**)	1	,535(**)	,816(**)
	Sig. (bilateral)	,000	,000		,000	,000
	N	74	74	74	74	74
Writing	Correlación de Pearson	,521(**)	,577(**)	,535(**)	1	,852(**)
	Sig. (bilateral)	,000	,000	,000		,000
	N	74	74	74	74	74
EOI	Correlación de Pearson	,642(**)	,800(**)	,816(**)	,852(**)	1
	Sig. (bilateral)	,000	,000	,000	,000	
	N	74	74	74	74	74

** La correlación es significativa al nivel 0,01 (bilateral).

Como podemos ver existe una correlación muy significativa entre los dos modelos de C-test y los tests de competencia de la EOI con $p < 0,01$, lo que demuestra la validez concurrente de cualquiera de los modelos de C-test. Los resultados también muestran que la correlación es mayor entre los C-tests y la prueba global de comprensión lectora (Reading) o el examen de competencia global de la EOI (EOI) y más baja entre los C-tests y el test de expresión escrita (Writing) o el test de comprensión oral (Listening). Estos resultados eran esperados, en cierta forma, ya que apoyan las dos teorías más comunes de los investigadores que dicen que el C-test mide principalmente la competencia global del idioma y la habilidad de la comprensión lectora. Alderson (2000), Connelly (1997), Daller y Phelan (2006), Raatz (1983), Klein-Braley (1985), y Kontra y kormos (2006).

Tabla 8.31. Correlaciones: C-test

		C-test	Lexis CTest	Function CTest	EOI	Tutor
C-test	Correlación de Pearson	1	,936(**)	,954(**)	,644(**)	,610(**)
	Sig. (bilateral)		,000	,000	,000	,000
	N	151	151	151	151	151
Lexis C-Test	Correlación de Pearson	,936(**)	1	,787(**)	,639(**)	,605(**)
	Sig. (bilateral)	,000		,000	,000	,000
	N	151	151	151	151	151
Function C-Test	Correlación de Pearson	,954(**)	,787(**)	1	,582(**)	,552(**)
	Sig. (bilateral)	,000	,000		,000	,000
	N	151	151	151	151	151
EOI [10]	Correlación de Pearson	,644(**)	,639(**)	,582(**)	1	,580(**)
	Sig. (bilateral)	,000	,000	,000		,000
	N	151	151	151	151	151
Tutor [11]	Correlación de Pearson	,610(**)	,605(**)	,552(**)	,580(**)	1
	Sig. (bilateral)	,000	,000	,000	,000	
	N	151	151	151	151	151

** La correlación es significativa al nivel 0,01 (bilateral).

La tabla 8.31 muestra que existe una correlación muy significativa entre el C-test y el test de competencia global de la EOI: con $r = 0{,}644$, y $p < 0{,}01$.

Si se eleva al cuadrado el valor del *coeficiente "r"* tendremos un *coeficiente de determinación* de 0.42, lo que significa que las dos series de notas se solapan en un 42% o, dicho de otro modo, que la varianza de una variable puede representar el 42% de la varianza de la otra variable. Es realmente interesante y significativo que las varianzas de las dos variables se solapen en un 42% si tenemos en cuenta que el C-test no forma parte de la batería de tests de la EOI. El coeficiente de validez podría haber sido incluso mayor, ya que todos los sujetos de nuestro estudio pertenecen al mismo nivel de competencia (Nivel Intermedio de la EOI) y de acuerdo con Brown (1988: 144) si las correlaciones se basan en una muestra con un nivel de competencia

[10] EOI: Test Global de la Escuela Oficial de Idiomas
[11] Tutor: Test de Control de Homogeneidad

del idioma relativamente homogéneo, los coeficientes de correlación van a ser bajos.

También podemos apreciar que la correlación del C-test de léxico con el test global de la EOI es un poco mayor que la del C-test funcional, lo que podría indicar que los términos léxicos miden mejor el nivel de competencia global de una lengua que los términos funcionales.

Las correlaciones más altas las encontramos entre el C-test global y los C-tests de léxico y funcional. Esto es lógico, ya que el C-test global está formado por los otros dos. En este caso el C-test funcional contribuye en mayor medida al C-test que el C-test de léxico.

El test del Tutor también correlaciona de forma muy significativa tanto con el C-test como con el test global de la EOI, lo que demuestra que era adecuado para actuar como test de control.

Tabla 8.32. Correlaciones: C-test

		C-test	RH	RMC	RCL	EOI
C-test	Correlación de Pearson	1	,258(**)	,430(**)	,562(**)	,644(**)
	Sig. (bilateral)		,001	,000	,000	,000
	N	151	151	151	151	151
RH [12]	Correlación de Pearson	,258(**)	1	,050	,440(**)	,566(**)
	Sig. (bilateral)	,001		,543	,000	,000
	N	151	151	151	151	151
RMC [13]	Correlación de Pearson	,430(**)	,050	1	,243(**)	,428(**)
	Sig. (bilateral)	,000	,543		,003	,000
	N	151	151	151	151	151
RCL [14]	Correlación de Pearson	,562(**)	,440(**)	,243(**)	1	,713(**)
	Sig. (bilateral)	,000	,000	,003		,000
	N	151	151	151	151	151
EOI	Correlación de Pearson	,644(**)	,566(**)	,428(**)	,713(**)	1
	Sig. (bilateral)	,000	,000	,000	,000	
	N	151	151	151	151	151

** La correlación es significativa al nivel 0,01 (bilateral).

[12] RH: Test de casar elementos o "Headings" incluido en la batería de tests de la comprensióm lectora.
[13] RMC: Test de elección múltiple incluido en la batería de tests de la comprensióm lectora.
[14] RCL: cloze test incluido en la batería de tests de la comprensióm lectora.

La tabla 8.32 nos da información sobre las correlaciones entre las diferentes tareas de la batería de tests con que se evalúa la comprensión lectora de la EOI. Puede verse que las correlaciones son significativas con $p < 0,01$. Como se esperaba, los coeficientes de correlación son más altos con el *cloze* (RCL) que con el resto de los tests, ya que el C-test es una clase de *cloze* y tanto el C-test como el *cloze* son tests basados en la misma teoría, que es la de redundancia reducida.

El *cloze* correlaciona ligeramente mejor que el C-test con el test global de la EOI. Esto es lógico, ya que el *cloze* forma parte de la batería de tests de dicho examen global y por lo tanto contribuye a su puntuación final.

Es sorprendente, sin embargo, que no exista una correlación significativa entre el test de elección múltiple y el test de *matching* o *headings* (RH) (casar extractos de pasajes con títulos). También es muy baja, aunque significativa, la correlación entre el test de elección múltiple y el *cloze.*

En cuanto a la contribución que cada test aporta al test global de la EOI, vemos que el *cloze*, aunque no estamos muy satisfechos con él, es el que contribuye en mayor medida al test de la EOI. Sin embargo, el tradicional test de elección múltiple es el que menor contribución realiza.

Según Alderson et al. (1995: 184), la razón de que exista esta correlación tan baja entre los diferentes tests que componen la batería de tests que evalúan la comprensión lectora puede deberse a que “they all measure something different and therefore contribute to the overall picture of language ability attempted by the test”. Si la correlación fuera realmente alta significaría que todos los tests estarían midiendo el mismo aspecto o la misma característica de la habilidad lectora con lo que podría eliminarse alguno de los tests que forman la batería.

La moderadamente alta correlación existente entre el C-test y la EOI indica que el solapamiento existente entre los dos constructos es más que

aceptable lo cual apoya la teoría de los investigadores que afirman que el C-test mide la competencia lingüística global de los candidatos (Raatz, 1983; Connelly, 1997; Klein-Braley, 1985; Eckes y Grotjanhn, 2006; Daller y Pelan, 2006; y Dörnyei y Katona, 1992).

Tabla 8.33. Correlaciones: C-test

		C-test	Reading	Listening	Writing	EOI
C-test	Correlación de Pearson	1	,580(**)	,520(**)	,441(**)	,644(**)
	Sig. (bilateral)		,000	,000	,000	,000
	N	151	151	151	151	151
Reading	Correlación de Pearson	,580(**)	1	,410(**)	,565(**)	,807(**)
	Sig. (bilateral)	,000		,000	,000	,000
	N	151	151	151	151	151
Listening	Correlación de Pearson	,520(**)	,410(**)	1	,462(**)	,805(**)
	Sig. (bilateral)	,000	,000		,000	,000
	N	151	151	151	151	151
Writing	Correlación de Pearson	,441(**)	,565(**)	,462(**)	1	,807(**)
	Sig. (bilateral)	,000	,000	,000		,000
	N	151	151	151	151	151
EOI	Correlación de Pearson	,644(**)	,807(**)	,805(**)	,807(**)	1
	Sig. (bilateral)	,000	,000	,000	,000	
	N	151	151	151	151	151

** La correlación es significativa al nivel 0,01 (bilateral).

Las correlaciones son también significativas al nivel 0,01 entre los tests que forman la batería del examen de la EOI, Reading, Listening, Writing, y el C-test, siendo las más altas las que se dan entre el Reading y el C-test y entre el Reading y el Writing y la más baja la que se da entre el Listening y el Reading. (Tabla 8.33).

Coleman (2002) también hizo una comparación entre los resultados del C-test con varios componentes del examen del “A-level”. Sus resultados fueron que aunque todas las correlaciones eran significativas, sin embargo, las correlaciones mas altas se daban entre el reading y el writing y entre el listening y el C-test. (Coleman 2002: 229).

Si la competencia global de la lengua es un constructo unitario, los procesos que forman parte de los conocimientos receptivos y productivos estarán íntimamente relacionados, por lo que las correlaciones entre los tests que usan diferente formato no deben sorprendernos.

La tabla 8.34 nos muestra que dentro de los tests que componen la prueba de comprensión lectora el que mejor correlaciona con el Reading es también el *cloze* y el que tiene una correlación más baja es el test de elección múltiple. Este último test es el que muestra las correlaciones más bajas con los otros dos componentes del test de Reading. Esto puede querer expresar que el test de elección múltiple mide algún aspecto de la habilidad lectora que difiere en gran medida con los aspectos que miden los otros dos tests, es decir, el *cloze* y el test de *headings*.

Tabla 8.34. Correlaciones: C-test

		C-test	EOI	Reading	RH	RMC	RCL
C-test	Correlación de Pearson	1	,641(**)	,580(**)	,256(**)	,424(**)	,562(**)
	Sig. (bilateral)		,000	,000	,002	,000	,000
	N	151	151	151	151	151	151
EOI	Correlación de Pearson	,641(**)	1	,807(**)	,567(**)	,430(**)	,711(**)
	Sig. (bilateral)	,000		,000	,000	,000	,000
	N	151	151	151	151	151	151
Reading	Correlación de Pearson	,580(**)	,807(**)	1	,728(**)	,597(**)	,789(**)
	Sig. (bilateral)	,000	,000		,000	,000	,000
	N	151	151	151	151	151	151
RH	Correlación de Pearson	,256(**)	,567(**)	,728(**)	1	,057	,440(**)
	Sig. (bilateral)	,002	,000	,000		,489	,000
	N	151	151	151	151	151	151
RMC	Correlación de Pearson	,424(**)	,430(**)	,597(**)	,057	1	,240(**)
	Sig. (bilateral)	,000	,000	,000	,489		,003
	N	151	151	151	151	151	151
RCL	Correlación de Pearson	,562(**)	,711(**)	,789(**)	,440(**)	,240(**)	1
	Sig. (bilateral)	,000	,000	,000	,000	,003	
	N	151	151	151	151	151	151

** La correlación es significativa al nivel 0,01 (bilateral).

Por otra parte, todos los tests que forman la prueba de comprensión lectora o Reading se correlacionan de forma significativa con el C-test, el Reading y el test global de la EOI. Estos datos confirman la idea inicial de

poder utilizar el formato de C-test como parte de los tests del examen de clasificación de la EOI, lo cual podría ser especialmente útil debido a la economía de este test.

8.7.3. Validez aparente del C-test

Recibimos información sobre la percepción que los candidatos tienen de la dificultad, lo apropiado y la relevancia del test a través de un cuestionario. Los alumnos expresaron sus opiniones sobre el C-test y los resultados, que coinciden con los hallados por Bradshaw (1990) en su estudio, fueron que el 72.2% de ellos encontraron el test difícil, aunque esto no se corresponde con los resultados obtenidos en los tests. Quizás ello se debió a que el test era completamente nuevo para ellos y siempre cuesta un poco adaptarse a un nuevo modelo de test. Las razones que esgrimieron para explicar por qué les resultaba difícil fueron que no tenían suficiente conocimiento de vocabulario, que no encontraban la palabra exacta, y que no podían escribir la palabra correcta ortográficamente.

Los alumnos no perciben que este test pueda medir el conocimiento gramatical de un idioma y se encuentran divididos en cuanto a fluidez y a conocimiento general se refiere. La mayoría de ellos coinciden en que el C-test mide principalmente vocabulario y ortografía. Sin embargo, también piensan que el test es adecuado y que puede ser un indicador de la competencia general de la lengua.

Cuando se les pregunta si les gustaría tener este test formando parte de la batería de tests de la EOI, el 50,3% contestan afirmativamente y el 49,7% contestan negativamente. De acuerdo con estos datos, podemos afirmar que el test no es aceptado completamente pero tampoco es rechazado al 100%. Debemos tener en mente que los métodos nuevos de examen, que además no

han sido practicados en clase, siempre originan un cierto rechazo o miedo a lo desconocido.

8.7.4. Consistencia interna del C-test

8.7.4.1. Fiabilidad de los dos modelos de C-test y del C-test global

La consistencia interna de los super-ítems del C-test se refiere a la medida en que los subtests que miden un aspecto particular del constructo de la lengua ínter correlacionan entre sí. La consistencia interna de los C-tests se ha calculado midiendo la consistencia entre los términos que lo componen con la fórmula Alfa de Cronbach, como aconsejaron los creadores del método (Klein-Braley 1997 y Raatz 1985:73). Cada texto es considerado como un super-ítem.

8.7.4.1.1. Fórmula Alfa de Cronbach

Las Tablas desde la 8.35 a la 8.43 muestran la fiabilidad de las dos formas equivalentes o paralelas del C-test y la fiabilidad del C-test global, la cual ha sido estimada administrando dos tests equivalentes (A y B) a un grupo de sujetos y calculando la correlación entre los resultados de los dos modelos. En este estudio los textos que hemos usado en los dos modelos son los mismos. Ya se ha demostrado que el C-test A y el C-test B son equivalentes y que los dos grupos de alumnos son homogéneos por lo que pueden ser considerados como si se tratara de un mismo grupo.

Los datos estadísticos (tablas 8.35, 8.38 y 8.41) muestran que el coeficiente de fiabilidad Alfa de Cronbach se puede considerar alto para todos

los tests (C-test A 0,878; C-test B 0, 872 y C-test Global 0, 828). Aunque no existe consenso en la literatura sobre el valor mínimo de una fiabilidad aceptable, Daller y Phelan (2006: 107) consideran que un valor Alfa de Cronbach a partir de 0,8 o incluso 0,6 es suficiente para temas de investigación.

Observamos que los valores de los coeficientes de fiabilidad de los modelos de C-test que hemos elaborado están en línea con los valores de los coeficientes Alfa de Cronbach de exámenes de reconocido prestigio y larga trayectoria. Así por ejemplo, la media del coeficiente Alfa de todos los modelos del test de comprensión de lectura académica del IELTS (International English Language Testing System) del año 2007 fue 0,86. El coeficiente de fiabilidad del MELAB (Michigan English Language Assessment Battery) del año 2000 estuvo entre 0,82 y 0,90 dependiendo del test. Finalmente, el Alfa de Cronbach del APIEL (Avanced Placement International English Language) que mide la competencia del inglés de hablantes no nativos y que se usa en muchas universidades americanas en lugar del TOEFL para decidir la admisión de alumnos a sus centros, varió entre 0,72 para el test de preguntas abiertas y 0,88 para el test de elección múltiple en el año 2002.

El valor Alfa de Cronbach está basado en las correlaciones entre los super-ítems o subtests, es decir, los cuatro textos que forman el C-test. Los datos estadísticos nos muestran que las correlaciones son bastante altas, especialmente las correlaciones entre los cuatro super-ítems que forman el C-test A y las correlaciones entre los cuatro super-ítems que forman el C-test B, lo cual indica que existe una considerable consistencia interna entre todos los elementos que componen los dos modelos de C-test que se han elaborado. Estos resultados son sorprendentes considerando que ninguno de los dos modelos de C-test se había pilotado y que tampoco se había realizado ninguna modificación en ninguno de los textos que los componen. Tampoco se realizó ningún análisis de los 100 ítems que componen el C-test. Además hay que tener en cuenta que los grupos de alumnos que hicieron los tests eran relativamente homogéneos, ya que todos ellos estaban estudiando el mismo

curso de la EOI, lo que origina que los coeficientes no sean tan altos como en el caso de que los grupos fueran más heterogéneos.

Tabla 8.35. Estadísticos de fiabilidad: C-test A

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,878	,886	4

Tabla 8.36. Matriz de correlaciones inter-elementos: C-test A[15]

	TtAPh	TtARe	TtAEn	TtAVi
TtAPh	1,000	,604	,600	,617
TtARe	,604	1,000	,762	,685
TtAEn	,600	,762	1,000	,693
TtAVi	,617	,685	,693	1,000

Tabla 8.37. Estadísticos total-elemento: C-test A

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento
TtAPh	47,0130	159,355	,673	,879
TtARe	51,1818	107,703	,793	,829
TtAEn	50,8312	119,247	,798	,819
TtAVi	53,4156	130,509	,757	,836

En el C-test A (tabla 8.36) todas las correlaciones entre los subtests son considerablemente altas sobresaliendo la correlación entre los subtests de "Relax and Live" y "Alternative Sources of Energy" que es 0,762. El subtest

[15] TtAPh: total physical exercise C-test A; TtARe: total relax and live C-test A; TtAEn: total alternative sources of energy C-test A; TtAVi total vitamins are vital C-test A.

"Physical Exercise" es el que tiene las correlaciones más bajas con el resto de los subtests aunque ninguna de ellas es inferior a 0,600.

La tabla 8.37 nos muestra que el elemento que más contribuye al C-test A es el subtest de "Alternative Sources of Energy" seguido del de "Relax and Live" que son los que mayor correlación muestran entre si. El subtest que menos contribuye es "Physical Exercise" que también coincide con el subtest que peor correlaciona con el resto de los subtests del C-test A.

Tabla 8.38. Estadísticos de fiabilidad: C-test B

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,872	,872	4

En el C-test B (tabla 8.39) vemos que las correlaciones entre los subtests han variado y ahora es el subtest "Physical Exercise" el que muestra una correlación más alta con el resto de elementos del test y el subtest "Relax and Live" el que muestra una correlación más baja

Tabla 8.39. Matriz de correlaciones inter-elementos: C-test B[16]

	TtBPh	TtBRe	TtBEn	TtBVi
TtBPh	1,000	,608	,654	,755
TtBRe	,608	1,000	,550	,540
TtBEn	,654	,550	1,000	,670
TtBVi	,755	,540	,670	1,000

[16] TtBPh: total physical exercise C-test B; TtBRe: total relax and live C-test B; TtBEn: total alternative sources of energy C-test B; TtBVi total vitamins are vital C-test B.

Tabla 8.40. Estadísticos total-elemento: C-test B

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento
TtBPh	53,3378	107,049	,793	,808
TtBRe	51,0270	128,492	,635	,870
TtBEn	51,2027	124,054	,720	,840
TtBVi	53,2973	110,047	,767	,819

Analizando la tabla 8.40 vemos que el subtest que más contribuye al C-test B es el de "Physical Exercise" seguido por el test de "Vitamins are Vital" y el que menos contribuye el de "Relax and Live". Esto quiere decir que prácticamente se ha invertido la situación respecto a lo que ocurría en el C-test A. Así pues, al variar el punto en el que se empieza a mutilar las palabras varían también las correlaciones entre los componentes del C-test y sus respectivas contribuciones aunque, como ya hemos demostrado con datos estadísticos, esto no influye en los resultados finales de los dos modelos de C-test.

Tanto los subtests del C-test A como los del C-test B están altamente correlacionados lo cual demuestra que todos los subtests miden la misma habilidad o constructo.

Tabla 8.41. Estadísticos de fiabilidad: C-test

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,828	,828	4

Tabla 8.42. Matriz de correlaciones inter-elementos: C-test

	Ttph	TtRe	TtEn	TtVi
TtPh	1,000	,370	,419	,460
TtRe	,370	1,000	,697	,638
TtEn	,419	,697	1,000	,694
TtVi	,460	,638	,694	1,000

Tabla 8.43. Estadísticos total-elemento: C-test[17]

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento
TtPh	50,1126	142,901	,469	,861
TtRe	51,1060	117,109	,688	,768
TtEn	51,0132	120,826	,746	,742
TtVi	53,3576	119,685	,734	,747

Finalmente, en el C-test global (tabla 8.42) los subtests que mejor correlacionan con el resto son “Alternative Sources of Energy” y “vitamins are vital” y el que tiene una correlación más baja es el de “Physical Exercise”. Vemos también (tabla 8.43) que los subtests “Alternative Sources of Energy” y “Vitamins are Vital” son los que mayor contribución aportan al C-test y “Physical Exercise” el que menos contribuye al mismo.

Podemos decir por lo tanto, que los subtests que más contribuyen al C-test son los que mejores correlaciones muestran con el resto de subtests que lo componen y por el contrario los que contribuyen en menor medida al C-test son los que tienen correlaciones más bajas con el resto de subtests que forman el C-test total.

[17] TtPh: total physical exercise C-test global; TtRe: total relax and live C-test global; TtEn: total alternative sources of energy C-test global; TtVi total vitamins are vital C-test global.

8.7.4.1.2. Análisis por mitades

Se estimó también la fiabilidad de los C-tests usando el método de las dos mitades o *Split-half.* En este caso se dividen los tests en dos mitades y se calcula la correlación entre los dos subtests. El coeficiente resultante se ajusta para calcular la fiabilidad del test completo. Para ello se utilizan las fórmulas de Spearman-Brown y la de Alfa (α) de Cronbach. Las Tablas 8.44, 8.45 y 8.46 muestran los resultados:

Tabla 8.44. Estadísticos de fiabilidad: C-test A

Alfa de Cronbach	Parte 1	Valor	,691
		N de elementos	2(a)
	Parte 2	Valor	,817
		N de elementos	2(b)
	N total de elementos		4
Correlación entre formas			,821
Coeficiente de Spearman-Brown	Longitud igual		,901
	Longitud desigual		,901
Dos mitades de Guttman			,900

a Los elementos son: TtAPh, TtARe.
b Los elementos son: TtAEn, TtAVi.

Las Tabla 8.44 nos muestra la consistencia interna entre los dos super-ítems (TtAPh y TtARe), que son los textos en los que se empezó a mutilar la segunda palabra de la segunda frase, y los super-ítems (TtAEn yTtAVi), que son los textos en los que se empezó a mutilar la tercera palabra de la segunda frase. Podemos observar que tanto los coeficientes de fiabilidad como los de correlación son bastante altos.

Tabla 8.45. Estadísticos de fiabilidad: C-test B

Alfa de Cronbach	Parte 1	Valor	,752
		N de elementos	2(a)
	Parte 2	Valor	,798
		N de elementos	2(b)
	N total de elementos		4
Correlación entre formas			,770
Coeficiente de Spearman-Brown	Longitud igual		,870
	Longitud desigual		,870
Dos mitades de Guttman			,870

a Los elementos son: TtBPh, TtBRe.
b Los elementos son: TtBEn, TtBVi.

La Tabla 8.45 nos muestra la consistencia interna entre los super-ítems del C-test B. En este Test se empezó suprimiendo la tercera palabra de la segunda frase en los dos primeros super-ítems (TtBPh y TtBRe), y la segunda palabra de la segunda frase en los dos últimos (TtBEn yTtBVi).

En los dos modelos de C-test observamos que el coeficiente de fiabilidad es ligeramente mayor para los dos últimos super-ítems. Puesto que la mutilación empieza con palabras diferentes en los dos modelos, podríamos llegar a la conclusión de que la consistencia del test puede verse afectada por la consistencia de los textos más que por las palabras que se mutilen.

Tabla 8.46. Estadísticos de fiabilidad: C-test

Alfa de Cronbach	Parte 1	Valor	,539
		N de elementos	2(a)
	Parte 2	Valor	,819
		N de elementos	2(b)
	N total de elementos		4
Correlación entre formas			,733
Coeficiente de Spearman-Brown	Longitud igual		,846
	Longitud desigual		,846
Dos mitades de Guttman			,845

a Los elementos son: Ttph, TtRe.
b Los elementos son: TtEn, TtVi.

La fiabilidad del C-test global (tabla 8.46) es más baja para los dos primeros super-ítems pero no así para los segundos. Aquí también se cumple, igual que en los dos modelos de C-test A y B, que el coeficiente de fiabilidad de los dos super-ítems “Alternative Sources of Energy” y “Vitamins are Vital” es bastante superior al de los otros dos componentes del C-test “Physical Exercise” y “Relax and Live”.

Capítulo 9

ANÁLISIS DEL CUESTIONARIO

9.1. Introducción

En el presente estudio intentamos determinar si el C-test es un método válido para recoger información que nos ayude a tomar decisiones sobre el progreso y competencia de los alumnos en una lengua. Para ello tenemos que tener en cuenta no sólo el conocimiento y las destrezas que un alumno posea sino también su actitud frente al idioma y al aprendizaje del mismo.

Analizando brevemente la situación actual del proceso de adquisición de una lengua, hay que señalar los cambios notables que se están produciendo en el área de lenguas extranjeras y su relativa importancia en el Sistema Educativo Español. Dichos cambios han sido motivados, en parte, por los avances de la psicolingüística y la sociolingüística y sus aplicaciones, así como por las demandas sociales sugeridas por los cambios políticos, económicos, sociales y laborales que han favorecido un desarrollo progresivo de la enseñanza de lenguas con fines profesionales y han determinado un enfoque centrado en las necesidades de los alumnos.

La Unión Europea en su Marco Común de Referencia para la enseñanza de idiomas sugiere a los diferentes Estados la necesidad de que sus ciudadanos hablen por lo menos dos lenguas extranjeras. Esta actitud en favor del dominio de más de una lengua extranjera por parte de los políticos europeos, junto con la aceptación por parte de los ciudadanos de la necesidad de obtener una formación inicial y continuada en lenguas extranjeras que les facilite la obtención de un puesto de trabajo, el desempeño de determinadas actividades o futuras promociones profesionales, hace que la enseñanza y evaluación de las lenguas extranjeras sean cada vez más exigentes tanto por parte de los alumnos como de los docentes.

Esta nueva situación ha requerido un planteamiento nuevo en la labor docente, en la de evaluación y en la de certificación. Hemos tenido que adaptar

programas o modificar el enfoque de los mismos para proporcionar a los alumnos en formación los instrumentos que los capaciten debidamente para desenvolverse y alcanzar sus propias metas en el futuro. En el caso de la EOI, esos instrumentos han sido los necesarios para propiciar la competencia comunicativa.

Nuestros alumnos son la mayoría adultos o adolescentes y deciden estudiar en las Escuelas de Idiomas bien por motivos laborales, ya que con un idioma pueden optar a un puesto mejor, o bien por motivos personales tales como: viajar, leer obras de literatura en su idioma original, libros y revistas científicas que no han sido publicados en Español o simplemente porque desean seguir aprendiendo a lo largo de su vida. La situación de nuestros alumnos no es pues la del estudiante propiamente dicho que dispone de mucho tiempo para el estudio y empieza con unos conocimientos de la asignatura similares a los del resto de sus compañeros de clase. La experiencia de nuestros alumnos con la lengua es muy variada. La mayoría de las personas más jóvenes han estudiado o están estudiando inglés en el colegio, otros, que tienen más edad o vienen de otros países, han estudiado otros idiomas y su contacto con el inglés es muy escaso.

El interés y la urgencia para aprender una lengua también son muy variados, ya que hay personas que van a ir a estudiar al extranjero o a trabajar con una multinacional donde van a usar el idioma con frecuencia y otros simplemente lo quieren para viajar con lo que no lo consideran algo prioritario en sus vidas. Lo que sí tienen en común es la escasez de tiempo disponible para dedicarlo al estudio, ya que la mayoría de ellos tienen responsabilidades familiares y laborales. Esta falta de tiempo y dedicación hace que muchos de estos alumnos adultos lleven estudiando una lengua de forma irregular e inconstante varios años sin que realmente avancen en sus conocimientos y como se dice coloquialmente “saben un poco de todo y mucho de nada”, lo cual les lleva a decepcionarse y a creerse incapaces de adquirir la competencia que necesitan o desearían en una lengua determinada.

9.2. Razones para usar el cuestionario

Debido a la complejidad del alumnado de la EOI decidimos analizar la situación de las personas que iban a tomar parte en nuestro estudio, tanto desde el punto de vista sociológico como de actitud y comportamiento frente al aprendizaje de la lengua inglesa, ya que estas variables podrían influir en los resultados de nuestra investigación. También estábamos interesados en conocer la percepción que los alumnos tenían del C-test, que era un método de examen totalmente nuevo para ellos. Nos interesaba especialmente saber si lo consideraban un método adecuado y completo así como su opinión sobre cuales podrían ser los conocimientos lingüísticos que esta nueva técnica podría medir. El instrumento elegido para la recogida de datos fue un cuestionario, que es el método más común para recoger datos en la investigación de una segunda lengua, debido a que, de acuerdo con Dörney (2003) un cuestionario es:

1. Fácil de construir.

2. Eficiente y económico tanto en términos de tiempo y esfuerzo empleado por el investigador como por los recursos financieros que se necesitan.

3. Extremadamente versátil. Es decir, se puede utilizar con éxito en multitud de situaciones, con personas muy variadas y con distintas finalidades. Por ello, la inmensa mayoría de proyectos de investigación en ciencias sociales y de comportamiento implican, en algún momento, recoger datos a través de un cuestionario

4. El único medio capaz de recoger gran cantidad de información que se puede procesar fácilmente y de forma inmediata, especialmente si usamos los modernos programas de ordenador de los que actualmente disponemos (Gillham 2000).

La Sociología, la Psicología social y la Psicometría están por delante de la Lingüística en esta clase de estudios. Brown denomina cuestionario a cualquier documento escrito que presente una serie de preguntas para ser contestadas bien respondiendo por escrito o bien seleccionando una respuesta de entre las propuestas.

> Questionnaires are any written instruments that present respondents with a series of questions or statements to which they are to react either by writing out their answers or selecting from among existing answers. (Brown, 2001: 6)

Los cuestionarios proporcionan información acerca de las personas que los realizan de "forma no evaluativa", ya que no miden sus respuestas ni de acuerdo con un criterio ni con respecto a las respuestas de otro grupo. Por lo tanto en los cuestionarios no existen respuestas correctas o incorrectas. Los datos obtenidos a través de un cuestionario son particularmente apropiados para el análisis estadístico cuantitativo, ya que la característica esencial de la investigación cuantitativa es que emplea categorías, puntos de vista y modelos que han sido definidos previamente por el investigador. (Gillham 2000). Estos datos además se pueden procesar fácilmente utilizando un programa de ordenador, lo cual es muy importante cuando se está trabajando e investigando a la vez y no se dispone de mucho tiempo.

9.3. ¿Que mide el cuestionario?

En general los cuestionarios nos permiten obtener tres tipos de datos: objetivos, de comportamiento y de actitud.

1. *Objetivos*. Las preguntas objetivas intentan recabar información acerca de los datos personales de los participantes tales como, nacionalidad, nivel de estudios, profesión, lengua materna, edad,

género, libro de texto que se utiliza en clase, tiempo que lleva estudiando la lengua, estado civil, etc.

2. *Comportamiento*. Las preguntas relacionadas con el comportamiento intentan saber lo que los participantes están haciendo o han hecho en el pasado. Se incluyen preguntas sobre el estilo de vida, hábitos, estrategias de aprendizaje que utilizan en un segundo idioma, frecuencia de uso de esas estrategias, etc.

3. *Actitud*. Las preguntas de actitud intentan obtener información sobre lo que los participantes piensan. En esta categoría se incluyen todas las preguntas relacionadas con las opiniones, creencias, actitudes, intereses y valores.

Los cuestionarios son muy utilizados para analizar las necesidades de los alumnos cuando se quiere diseñar o programar un curso y decidir con antelación el enfoque didáctico, los temas, las actividades o las destrezas que interesa practicar en el mismo para sacarle el máximo rendimiento posible. También se utiliza en las empresas, cuando se hacen estudios de mercado, para decidir cual será el producto que tendrá más aceptación por los futuros clientes, qué características deberá presentar, qué precio deberá costar, cual será la época propicia para lanzarlo al mercado, etc.

El análisis de necesidades se ha considerado como un aspecto característico de los cursos de inglés, ya sea general o específico. Según West (1984), estos análisis se comenzaron a principios de los años setenta y se centraron principalmente en los cursos de inglés para fines específicos (ESP o English for Specific Purposes) y cursos ocupacionales (EOP o English for Ocupational Purposes), (Richterich 1973). A finales de los años setenta se incluyó también el inglés académico (EAP o English for Academic Purposes) dentro de estos análisis y se estudiaba la situación laboral y profesional que los alumnos se iban a encontrar cuando acabaran el curso. De esta forma, una vez se conocían las necesidades, se intentaba diseñar el curso de acuerdo a ellas (Mackay 1978).

En los años ochenta el análisis de necesidades se fue ampliando incluyendo aspectos tales como las estrategias de aprendizaje, los análisis de deficiencias y de medios, y los materiales de enseñanza. Además, se empezaron a realizar análisis de necesidades también en áreas como el inglés general (GE). (Tarone y Yule 1989; Allwright & Allwright 1977; y Allwright 1982).

A comienzo de los años noventa se comenzó con los análisis integrados o asistidos por ordenador y se incluyó el área de selección de materiales de enseñanza. Los análisis de necesidades se centraban principalmente en los cursos de inglés para fines específicos.

Aunque como vemos el uso del cuestionario para realizar análisis de necesidades tiene ya un largo recorrido, sin embargo, no parece existir consenso a la hora de definir qué es una necesidad lingüística. De acuerdo con Richterich (1983: 2): “the very concept of language need has never been clearly defined and remains at best ambiguous”.

Hutchinson y Waters (1987), aunque también opinan que el término *necesidad* se presta a varias interpretaciones, distinguen entre tres tipos de necesidades:

1. Las demandas o necesidades objetivas, es decir, “the type of need determined by the demands of the target situation, that is, what the learner has to know in order to function effectively in the target situation”.

2. Los deseos o necesidades subjetivas del alumno, las cuales se han definido como “what the learners want or feel they need”.

3. Las deficiencias o carencias de los alumnos que sería la diferencia entre lo que el alumno sabe y lo que necesita saber para satisfacer sus necesidades subjetivas.

Nuestro caso no es ninguno de estos y lo que queremos conseguir con el cuestionario es obtener un perfil de las características biográficas de los participantes y conocer los intereses, las expectativas y la actitud de los estudiantes ante el aprendizaje de la lengua. Este perfil nos permitirá, por una parte, identificar y analizar la motivación, las necesidades funcionales y los contextos comunicativos de uso real de los alumnos, y por otra parte, nos facilitará la identificación y el análisis de las preferencias, los estilos y las necesidades de aprendizaje de los participantes en la investigación.

El cuestionario que se elaboró es un cuestionario escrito y cerrado. Esto quiere decir que los alumnos reciben tanto las preguntas como las respuestas escritas y ellos sólo tienen que seleccionar la respuesta específica más adecuada para expresar sus opiniones o sentimientos. En el cuestionario (ver apéndice 3) se distinguen tres apartados:

1. Datos sociométricos u objetivos.
2. Datos referentes a su actitud, comportamiento y hábitos frente al aprendizaje de la lengua.
3. Percepción que los alumnos tienen del C-test.

La opinión que los alumnos tengan del C-test nos informará de la Validez Aparente del C-test.

9.4. Procedimiento

El cuestionario se realizó en la misma clase después de haber terminado el C-test y sin tiempo límite para contestar a las preguntas. Los profesores que habían vigilado el examen se encargaron también de administrar el cuestionario.

A los alumnos se les prometió explícitamente confidencialidad con todas las respuestas y datos personales y se les explicó que se les pedía el nombre para poder asociar los resultados del C-test con sus respectivas respuestas al cuestionario. También se les explicó que el nombre nunca aparecería en el estudio, ya que éste iba a ser asociado a un número que sería lo que aparecería en el trabajo de investigación. Así mismo, se les dijo en qué consistía el estudio haciendo hincapié en que no había respuestas correctas o incorrectas y que era importante que las contestaciones estuvieran lo más cerca posible a la realidad.

Una vez recogidos los cuestionarios, se introdujeron las respuestas en la base de datos de un programa de ordenador para ser analizadas.

9.5. Análisis de resultados

9.5.1. Datos sociométricos: descripción de los participantes

Este grupo de 151 estudiantes se caracteriza por estar formado mayoritariamente (74,2%) por mujeres y por personas universitarias o profesionales con estudios universitarios. El 82,8% de los alumnos poseen titulaciones universitarias y la mayoría de ellos se encuentran insertados en el mercado laboral (tablas 9.1 y 9.2).

Tabla 9.1. Sexo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Varón	39	25,8	25,8	25,8
	Mujer	112	74,2	74,2	100,0
	Total	151	100,0	100,0	

Tabla 9.2. Estudios

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Primarios	5	3,3	3,3	3,3
	Bachillerato	21	13,9	13,9	17,2
	universitarios	125	82,8	82,8	100,0
	Total	151	100,0	100,0	

En cuanto a la edad, el 59,6% tienen menos de 30 años y el 40,4% tienen una edad igual o superior a 30 años.

9.5.2. Actitud y comportamiento de los participantes ante el aprendizaje de la lengua

Se considera que el alumno es un agente activo que contribuye a su formación y al aprendizaje de la lengua con actuaciones concretas y actitudes positivas. La actitud de la persona que está aprendiendo un idioma está íntimamente relacionada con la motivación.

Para analizar la actitud de los alumnos ante el aprendizaje de la lengua nos hemos basado en la *Teoría de la Autodeterminación* de Decy y Ryan (2000). Estos autores analizan cómo la motivación favorece el proceso de enseñanza-aprendizaje. En concreto:

1. Favorece una actitud activa y comprometida.
2. Coadyuva al mantenimiento del esfuerzo y a la dedicación necesaria para el aprendizaje.
3. Promueve la asimilación y el desarrollo de las habilidades.
4. Favorece el sentimiento de competencia.

Decy y Ryan (2000) consideran la motivación como la causa que impulsa a una persona a actuar y promueve la autonomía, el sentimiento de competencia, la curiosidad y el deseo de asumir desafíos. La motivación no es lineal sino que está vinculada con la actividad concreta que el alumno tiene que realizar. Cuanto mayor es la motivación mayor es la implicación, mejor el rendimiento, menor el fracaso y mayor la calidad de aprendizaje.

De acuerdo con El Consejo de Europa (2001):

> El uso de la lengua – que incluye el aprendizaje – comprende las acciones que realizan las personas que, como individuos y como agentes sociales, desarrollan una serie de competencias, tanto generales como competencias comunicativas, en particular.

Las preguntas que se han elaborado para ver la actitud de los alumnos frente al aprendizaje de la lengua están relacionadas con su comportamiento y las actividades que realizan fuera de la clase que pueden contribuir a mejorar el conocimiento que poseen de la lengua.

Al analizar los resultados vemos que un dato de gran relevancia es el elevado porcentaje de alumnos que lleva estudiando inglés más de cinco años (el 75,5%), el 47% de los cuales lleva estudiando inglés durante más de 10 años.

Tabla 9.3: Uso de la lengua inglesa

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No	45	29,8	29,8	29,8
	Sí	106	70,2	70,2	100,0
	Total	151	100,0	100,0	

La mayoría de los alumnos dicen usar el inglés fuera de clase (tabla 9.3), aunque solamente un 34,4% dicen hacerlo con frecuencia.

Tabla 9.4: Personas que hablan inglés fuera de clase

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No	37	24,5	24,5	24,5
	Sí	114	75,5	75,5	100,0
	Total	151	100,0	100,0	

El 75,5% también afirman hablar inglés fuera de clase (tabla 9.4). Sin embargo, la frecuencia no es muy alta, ya que sólo el 20,5% de ellos lo hace frecuentemente, el 39,1% dice hablar en inglés de vez en cuando y el resto solamente en vacaciones.

El 48,3% de los alumnos dicen asistir hasta el 90% de las clases y otro 37,7% afirman asistir hasta el 75% de las clases. Los que aseguran hacer los deberes entre el 50% y el 75% de los días alcanzan un 61%.

Del 74,2% de las personas que aseguran leer en inglés, el 31,8% dice hacerlo frecuentemente y el 26,5% esporádicamente. Algo parecido ocurre con las personas que aseguran ver películas en inglés (el 81,5%), de las cuales el 48,3% afirman que las ven de vez en cuando, el 21,9% lo hace mensualmente y el 11,3% frecuentemente.

El 59,6% de las personas que escriben en inglés y el 51% de las que hablan en inglés dicen intentar pensar en inglés y no en español. El resto aseguran pensar siempre en español o siempre en inglés.

Según Dörney (2003), uno de los problemas que nos encontramos cuando intentamos medir actitudes, creencias, opiniones, valores, aspiraciones, expectativas y otras variables personales con cuestionarios, es que diferentes personas pueden interpretar las preguntas de forma diferente dependiendo de lo que cada persona entienda por las opciones: “frecuentemente”, “de vez en cuando”, “bastante bien”, “muy adecuado”, etc. Oppenheim también considera

esto un verdadero problema cuando afirma que todavía no se sabe por qué preguntas muy similares dan lugar a respuestas tan distintas.

> When we sometimes despair about the use of language as a tool for measuring or at least uncovering awareness, attitude, precepts and belief systems, it is mainly because we don not yet know *why* questions that look so similar actually produce such very different sets of results, or how we can predict contextual effects on a question, or in what ways we can ensure that respondents will all use the same frame of reference in answering an attitude question. (Oppenheim, 1992:149)

9.5.3. Percepción del C-test por los participantes

Las investigaciones empíricas han demostrado que el constructo del C-test es bastante complejo. Sin duda comprende el conocimiento general de la lengua (conocimiento ortográfico, morfológico, léxico y sintáctico) así como el conocimiento de propiedades del texto como son la coherencia y la cohesión. Además comprende también la habilidad de procesar la lengua a todos los niveles desde una letra individual hasta el texto completo.

Según Sigott la medida en que un pasaje de C-test nos puede dar información sobre el conocimiento de un candidato de todos estos aspectos no es fácilmente predecible, ya que depende de la dificultad del pasaje y de la habilidad del candidato.

> The extent to which these different aspects of the construct are tapped by the individual C-test passage is likely to be a function of person ability and passage difficulty, and therefore not easily predictable. (Sigott, 2004: 200)

Al ser un nuevo tipo de examen en la EOI, nos interesaba saber la opinión de los alumnos sobre el mismo por lo que en el cuestionario incluimos

unas preguntas específicas para conocer lo que los alumnos creían que media el C-test, si era una prueba adecuada y completa, si la consideraban difícil, y si eran favorables a la inclusión del C-test en la batería de tests de la EOI. Las respuestas que se obtendrían en este apartado nos ayudarían a determinar la validez aparente del C-test.

9.5.4. El C-test y el conocimiento técnico de la lengua

En el cuestionario se diseñaron preguntas para recabar información de hasta que punto los alumnos opinaban que el C-test servia para medir algunos aspectos del conocimiento técnico de la lengua, tales como la gramática, la ortografía, el conocimiento global del idioma, la fluidez o el vocabulario.

Al analizar las respuestas de los alumnos se observaron ciertas incoherencias entre lo que piensan y los datos reales. Por ejemplo, un amplio número de alumnos, el 72,2%, considera que el test era difícil. Esta opinión, sin embargo, no se ajusta a la realidad, ya que al analizar los resultados del C-test vemos que según los datos estadísticos (ver las medias en la tabla 8.1) los porcentajes de aciertos obtenidos en los subtests que forman el C-test oscilan entre el 60,7% y el 73,6%, muy por encima del 50% ó 60% recomendado por los creadores del test, y las medias del C-test A y el C-test B son 67,48 y 69,68 respectivamente sobre un total de 100 puntos, también por encima de los valores recomendados (tablas 8.2 y 8.3)

Entre el 85% y el 90% de los alumnos opinan que el grado en que el C-test mide la gramática y la ortografía estaría entre “bastante” o “mucho”. Sin embargo, a la hora de determinar el grado en que el C-test mide el conocimiento general de la lengua un 88% de los que han realizado el examen se decantan por las opciones “poco” (41,1%) o “bastante” (47,7%), prácticamente a partes iguales.

En cuanto a la fluidez y al léxico vemos que casi la mitad de los sujetos, el 46,4% exactamente, consideran que el C-test mide “mucho” la fluidez del idioma mientras que cuando se les pregunta por el léxico un 49,0%, también casi la mitad de los participantes en el estudio, piensan que el C-test mide “bastante” el conocimiento de léxico de una persona.

Hemos querido estudiar si existía alguna relación entre todos estos aspectos que definen el conocimiento técnico de la lengua (tabla 9.5). Para ello hemos calculado los coeficientes de correlación Tau_b de Kendall y Rho de Spearman. Los dos coeficientes ofrecen valores muy similares, aunque los del primer coeficiente son un poco más bajos que los del segundo.

Tabla 9.5. correlaciones gramaticales y del conocimiento técnico de la lengua

			Gramática	Spelling	Conocimiento	Fluidez	Léxico
Tau_b de Kendall	Gramática	Coeficiente de correlación	1,000	,062	,180(*)	,168(*)	-,070
		Sig. (bilateral)	.	,408	,017	,023	,345
		N	151	151	151	151	151
	Spelling	Coeficiente de correlación	,062	1,000	,130	,161(*)	,360(**)
		Sig. (bilateral)	,408	.	,088	,032	,000
		N	151	151	151	151	151
	Conocimiento	Coeficiente de correlación	,180(*)	,130	1,000	,257(**)	,136
		Sig. (bilateral)	,017	,088	.	,001	,070
		N	151	151	151	151	151
	Fluidez	Coeficiente de correlación	,168(*)	,161(*)	,257(**)	1,000	,017
		Sig. (bilateral)	,023	,032	,001	.	,823
		N	151	151	151	151	151
	Léxico	Coeficiente de correlación	-,070	,360(**)	,136	,017	1,000
		Sig. (bilateral)	,345	,000	,070	,823	.
		N	151	151	151	151	151
Rho de Spearman	Gramática	Coeficiente de correlación	1,000	,068	,193(*)	,183(*)	-,080
		Sig. (bilateral)	.	,410	,017	,024	,331
		N	151	151	151	151	151
	Spelling	Coeficiente de correlación	,068	1,000	,139	,174(*)	,381(**)
		Sig. (bilateral)	,410	.	,089	,032	,000
		N	151	151	151	151	151
	Conocimiento	Coeficiente de correlación	,193(*)	,139	1,000	,278(**)	,147
		Sig. (bilateral)	,017	,089	.	,001	,072
		N	151	151	151	151	151
	Fluidez	Coeficiente de correlación	,183(*)	,174(*)	,278(**)	1,000	,018
		Sig. (bilateral)	,024	,032	,001	.	,828
		N	151	151	151	151	151
	Léxico	Coeficiente de correlación	-,080	,381(**)	,147	,018	1,000
		Sig. (bilateral)	,331	,000	,072	,828	.
		N	151	151	151	151	151

De acuerdo con los coeficientes de correlación mostrados en la Tabla 9.5, podemos afirmar que las respuestas que dan los participantes en el proyecto de investigación son bastante aleatorias. No parece existir mucha consistencia a la hora de definir los distintos aspectos que forman parte del conocimiento técnico de la lengua. Solamente las correlaciones entre el léxico y el “spelling” y entre el conocimiento y la fluidez son bastante significativas, aunque con valores no muy altos. También son significativas, pero muy bajas,

las correlaciones entre el conocimiento y la gramática y entre la fluidez y la gramática o el "spelling".

Estos datos pueden significar o que los alumnos no tienen muy claro los conceptos que les hemos pedido que valoren, o bien que no están seguros de lo que realmente mide el C-test.

9.5.5. Valoración del C-test

Las cuatro últimas preguntas del cuestionario se centran en recabar información de los alumnos acerca del diseño del C-test. Se les preguntó si creían que el modelo de examen era adecuado, completo, reflejaba bien sus conocimientos de inglés, y si les gustaría que el C-test formara parte del examen de comprensión lectora o lo que es lo mismo si consideran que el C-test se podría utilizar como complemento a la batería de tests que forman el examen de Reading. A esta última pregunta 76 alumnos, de los 151 que hicieron el test, contestaron que si y 75 contestaron que no. Es decir, que la opinión está totalmente dividida.

El 43,7% de las personas consideran que el test es bastante adecuado y el 32,5% que es adecuado (tabla 9.6). Sin embargo, hasta un 43,7% de los alumnos piensan que el C-test es un test poco completo y el 39,1% que es bastante completo.

Tabla 9.6. Adecuación

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Poco	36	23,8	23,8	23,8
	Bastante	66	43,7	43,7	67,5
	Adecuada	49	32,5	32,5	100,0
	Total	151	100,0	100,0	

Tabla 9.7. Indicador

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	malo	54	35,8	35,8	35,8
	Aceptable	64	42,4	42,4	78,1
	Bueno	33	21,9	21,9	100,0
	Total	151	100,0	100,0	

En cuanto a que el C-test pueda ser un indicador, es decir, pueda reflejar sus conocimientos de la lengua, casi la mitad de las personas, el 42,4%, piensan que el test puede ser un indicador aceptable y un 21,9% que puede ser un buen indicador (tabla 9.7).

Hemos calculado de nuevo los coeficientes de correlación Tau_b de Kendall y Rho de Spearman para las preguntas referidas al diseño del C-test, y observamos que existen correlaciones significativas y con valores de los coeficientes aceptables entre los ítems que recogen la opinión de los alumnos sobre la técnica de examen del C-test (tabla 9.8).

Tabla 9.8. Correlaciones

			Adecuación	Completo	Indicador
Tau_b de Kendall	Adecuación	Coeficiente de correlación	1,000	,627(**)	,512(**)
		Sig. (bilateral)	.	,000	,000
		N	151	151	151
	Completo	Coeficiente de correlación	,627(**)	1,000	,609(**)
		Sig. (bilateral)	,000	.	,000
		N	151	151	151
	Indicador	Coeficiente de correlación	,512(**)	,609(**)	1,000
		Sig. (bilateral)	,000	,000	.
		N	151	151	151
Rho de Spearman	Adecuación	Coeficiente de correlación	1,000	,671(**)	,562(**)
		Sig. (bilateral)	.	,000	,000
		N	151	151	151
	Completo	Coeficiente de correlación	,671(**)	1,000	,655(**)
		Sig. (bilateral)	,000	.	,000
		N	151	151	151
	Indicador	Coeficiente de correlación	,562(**)	,655(**)	1,000
		Sig. (bilateral)	,000	,000	.
		N	151	151	151

** La correlación es significativa al nivel 0,01 (bilateral).

La encuesta permite constatar que no existe ni una aceptación ni un rechazo totalmente claros del C-test. Este resultado es acorde con lo esperado considerando que este modelo de test es completamente nuevo para ellos y siempre hay una cierta reticencia a aceptar nuevos métodos de examen que no han sido practicados y que por lo tanto pueden crear inseguridad.

9.5.6. Relaciones entre las variables "uso del inglés" y "hablar"

Queremos conocer si existe alguna relación entre algunos ítems del cuestionario o por el contrario no hay relación alguna. Nos centramos para empezar, en observar si existía alguna relación entre las personas que decían

usar normalmente el idioma inglés y las que afirmaban hablar con alguna persona en inglés. Según la información proporcionada en este apartado por las tres tablas siguientes observamos que la tabla 9.9 nos presenta un χ^2 de 24,533 con una sig. = 0,000 y un grado de libertad de 1. Esto indica que hay una asociación positiva significativa entre las dos variables: "uso" y "hablar".

Tabla 9.9. Pruebas de χ^2

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	24,533(b)	1	,000		
Corrección por continuidad(a)	22,527	1	,000		
Razón de verosimilitudes	23,050	1	,000		
Estadístico exacto de Fisher				,000	,000
Asociación lineal por lineal	24,370	1	,000		
N de casos válidos	151				

a Calculado sólo para una tabla de 2x2.
b 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 11,03.

En la tabla 9.10, observamos que la casilla del "Sí" de los residuos (con un valor de 5,0) en la que se cruzan las dos variables contribuye a explicar significativamente el χ^2, el resto de las casillas, al tratarse de una tabla de contingencia de 2 x 2, son redundantes.

Tabla 9.10. de contingencia Uso * Hablar

			Hablar		Total
			No	Sí	
Uso	No	Recuento	23	22	45
		Residuos corregidos	5,0	-5,0	
	Sí	Recuento	14	92	106
		Residuos corregidos	-5,0	5,0	
Total		Recuento	37	114	151

9.5.6.1. Magnitud de la asociación

La magnitud de la asociación se observa en la tabla (9.11). Tanto el estadístico Phi como de la V de Cramer, que coinciden porque se trata de una tabla de contingencia de 2 x 2, nos dan una cifra similar. La magnitud del grado de asociación alcanza un valor de 0,403 que puede considerarse razonable.

Tabla 9.11. Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,403	,000
	V de Cramer	,403	,000
N de casos válidos		151	

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

9.5.7. Relación entre "conocimiento general de la lengua" y "fluidez"

Entre las preguntas que se hicieron, para saber si los alumnos pensaban que el C-test reflejaba el conocimiento técnico de la lengua, estaba la de si el test media la fluidez y el conocimiento general de la lengua. La tabla 9.12 nos presenta un χ^2 de 14,993 con una sig. = 0,005 y con cuatro grados de libertad; lo cual indica que hay una asociación positiva significativa entre las dos variables: "conocimiento" y "fluidez".

Tabla 9.12. Pruebas de χ^2

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	14,993(a)	4	,005
Razón de verosimilitudes	15,612	4	,004
Asociación lineal por lineal	10,764	1	,001
N de casos válidos	151		

a 1 casillas (11,1%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,72.

En la tabla 9.13 vemos que hay cuatro casillas de residuos con valores de 3,0; -2,9; -2,7 y 2,1 que contribuyen significativamente al χ^2. La primera y la última muestran una asociación positiva significativa y la 2ª y la 3ª muestran, por el contrario una asociación negativa significativa.

Tabla 9.13. de contingencia Conocimiento * Fluidez

			Fluidez			
			Poco	Bastante	Mucho	Total
Conocimiento	Poco	Recuento	21	21	20	62
		Residuos corregidos	3,0	,5	-2,9	
	Bastante	Recuento	9	25	38	72
		Residuos corregidos	-2,7	,7	1,5	
	Mucho	Recuento	3	2	12	17
		Residuos corregidos	-,4	-1,9	2,1	
Total		Recuento	33	48	70	151

9.5.7.1. Magnitud de la asociación

El tamaño de la asociación lo observamos en la siguiente tabla:

Tabla 9.14. Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,315	,005
	V de Cramer	,223	,005
N de casos válidos		151	

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

De los valores que nos muestra la tabla 9.14, al tratarse de una tabla de contingencia de 3 x 3, para valorar la magnitud de la asociación sólo se tiene en cuenta la V de Cramer y no el estadístico Phi. La magnitud del grado de

asociación alcanza un valor de 0,223 que puede considerarse un valor medio bajo.

9.5.8. Relación entre las variables "indicador" y "completo"

En el cuestionario se diseñaron cuatro preguntas para saber cual seria la aceptación del C-test en el caso de que se decidiera incluirlo en la batería de tests de la EOI. Entre las respuestas obtenidas vamos a analizar dos de ellas que hemos denominado como: "indicador" y "completo".

La variable "indicador" se refiere a la pregunta de si creen que el C-test reflejará bien sus conocimientos de inglés, es decir, si el C-test puede actuar como indicador de los conocimientos de inglés de una persona. En cuanto a la variable "completo", ésta corresponde a la pregunta de si creen que el C-test es un examen completo.

Tabla 9.15. Pruebas de χ^2

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	78,103(a)	4	,000
Razón de verosimilitudes	81,386	4	,000
Asociación lineal por lineal	64,945	1	,000
N de casos válidos	151		

a 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 5,68.

La tabla 9.15 nos presenta un χ^2 de 78,10 con una sig. = 0,000 y con 4 grados de libertad, lo cual indica que hay una asociación positiva significativa entre las dos variables: "indicador" y "completo".

Tabla 9.16. de contingencia Completo * Indicador

			Indicador			
			malo	Aceptable	Bueno	Total
Completo	Poco	Recuento	44	20	2	66
		Residuos corregidos	7,0	-2,6	-4,9	
	Bastante	Recuento	10	36	13	59
		Residuos corregidos	-3,9	3,7	,0	
	Completo	Recuento	0	8	18	26
		Residuos corregidos	-4,2	-1,3	6,4	
Total		Recuento	54	64	33	151

La tabla 9.16 nos muestra que las casillas de residuos que contribuyen significativamente al valor del χ^2 mostrando una asociación positiva son las que aparecen en la diagonal, con valores de 7,0; 3,7; y 6,4; confirmando el sentido esperado.

9.5.8.1. Magnitud de la asociación

El tamaño de la asociación lo observamos en la siguiente tabla:

Tabla 9.17. Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,719	,000
	V de Cramer	,509	,000
N de casos válidos		151	

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

La tabla 9.17 nos muestra el grado de asociación entre “completo” e “indicador”. Al tratarse también de una tabla de contingencia de 3 x 3, sólo se tiene en cuenta la V de Cramer para valorar la magnitud de la asociación, que en este caso presenta un valor de 0,509 bastante más alta que en los dos casos anteriores y que puede considerarse muy buena.

9.5.9. Relación entre los resultados del C-test y algunos ítems del cuestionario

En concreto se analiza si existe alguna relación entre los resultados que los alumnos han obtenido en el C-test y el uso, la lectura, la edad o el sexo.

Tabla 9.18, Estadísticos de grupo

	Uso	N	Media	Desviación típ.	Error típ. de la media
C-Test	No	45	64,2000	12,99056	1,93652
	Sí	106	70,3679	14,77922	1,43548

En la tabla 9.18 vemos que la mayoría de los alumnos (106 de 151) usan normalmente el idioma inglés y también que estos han obtenido una media más alta en los resultados del C-test que aquellos alumnos que no lo utilizan. Como nos interesa saber si esas diferencias entre las medias de los resultados de los alumnos que usan el inglés y los que no lo usan son significativas, hallamos la tabla del t-test aplicada a medir esta diferencia.

Tabla 9.19. Prueba de muestras independientes

		Prueba T para la igualdad de medias						
							95% Intervalo de confianza para la diferencia	
		t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Superior	Inferior
C-Test	Se han asumido varianzas iguales	-2,429	149	,016	-6,16792	2,53972	-11,18645	-1,14940
	No se han asumido varianzas iguales	-2,559	93,775	,012	-6,16792	2,41054	-10,95426	-1,38159

En esta tabla vemos que el valor del t-test cuando se han asumido varianzas iguales es de -2,429 con un gl = 149 y p = 0,016 y de -2,559 con un gl = 93,775 y p = 0,012 cuando no se han asumido varianzas iguales.

Estos datos nos confirman que existen diferencias significativas entre las medias de los resultados del C-test de los alumnos que usan y los que no usan normalmente el inglés, obteniendo mejores resultados en el test los alumnos que utilizan el idioma inglés.

9.5.10. Relación entre los resultados del C-test y la lectura

A continuación estudiaremos si existe alguna relación entre los resultados de C-test de los alumnos que leen en inglés frente a los que no lo hacen.

En la tabla 9.20 vemos que 112 personas de las 151 que han formado parte de este estudio leen en inglés y que sus resultados en el C-test son mejores que los de las personas que no lo hacen.

Tabla 9.20. Estadísticos de grupo

	Lectura	N	Media	Desviación típ.	Error típ. de la media
C-Test	No	39	60,9231	13,66206	2,18768
	Sí	112	71,1786	13,89356	1,31282

En la tabla 9.21 vemos que las diferencias entre las medias de los alumnos que leen en inglés y las de los que no lo hacen son significativas tanto si se asumen como si no se asumen varianzas iguales, con valores de t = -3,987 en el primer caso y t = -4,020 en el segundo y un valor de significacion: sig = 0,000 en ambos casos

Tabla 9.21. Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bi)	Diferencia de medias	Error típ. de la diferencia	Superior	Inferior
C-Test	Se han asumido varianzas iguales	,009	,924	-3,987	149	,000	-10,25549	2,57231	-15,33841	-5,17258
	No se han asumido varianzas iguales			-4,020	67,308	,000	-10,25549	2,55136	-15,34760	-5,16339

9.5.11. Relación entre los resultados del C-test, el test de la EOI y la edad

Dado que la enseñanza de idiomas en la EOI está dirigida a personas adultas, nos encontramos conque los alumnos tienen edades muy diferentes. En una misma clase podemos tener desde personas adolescentes hasta personas jubiladas. Como ya vimos en los datos sociométricos, el 40,4% de los

alumnos tienen una edad igual o superior a 30 años y el 59,6% son menores de esa edad.

Uno de los debates más vivos en la investigación del Aprendizaje de una Segunda Lengua (SLA: Second Language Acquisition) está relacionado con encontrar algún fundamento científico que demuestre la superioridad de las personas jóvenes sobre la de las personas mayores a la hora de adquirir una segunda lengua.

Desde que Penfield y Roberts (1959) propusieron que el cerebro gradualmente pierde plasticidad y Lenneberg (1967) formuló la Hipótesis del Periodo Crítico (CPH, Critical Period hipothesis), los investigadores han discutido a favor y en contra de la existencia de un periodo crítico (CP) para aprender una lengua. Durante los años 1960 y 1970 el debate estaba centrado en determinar si los niños realmente tenían alguna ventaja, ya que a veces estudios empíricos sugerían lo contrario. Sin embargo, desde que Krashen et al. (1979) publicaron su artículo, en el que se distinguía entre la velocidad inicial y la consecución final del aprendizaje de una lengua, quedaba claro que aunque las personas de más edad pueden tener una ventaja inicial en la adquisición de aspectos morfosintácticos de una L2, la consecución de los últimos niveles correlaciona negativamente con el aumento de la edad en que se empieza a estudiar una lengua. Con este estudio lo ya observado por las personas no expertas en la materia de que los niños aprenden mejor un idioma que las personas mayores recibía apoyo científico. Por lo tanto, en los años 1980 y 1990 el debate se centró en determinar las causas de esa ventaja a favor de los jóvenes estudiantes. Mientras unos se centraron en estudiar si existían causa biológicas o psicológicas otros, tales como Johnson y Newport (1989), confiaban en los datos que obtenían sobre la consecución de los últimos niveles.

A pesar de los numerosos estudios empíricos que se han publicado, no se ha avanzado mucho desde el estudio de Krashen *et al*. (1979). Solamente se puede afirmar que a veces las personas que empiezan más tarde a estudiar una segunda legua tienen mejores resultados en las primeras etapas de

adquisición morfosintáctica que las más jóvenes. Sin embargo, las personas más jóvenes generalmente adquieren niveles de competencia semejante a las personas nativas, mientras que, no hay evidencia de que las personas que empiezan a estudiar un idioma cuando son adultas puedan alcanzar un nivel de competencia similar a los nativos.

Algunos autores están a favor de la existencia de un periodo crítico después del cual no se puede aprender un idioma. Otros autores, por ejemplo. Birdsong (1999), apuestan por la posibilidad de que exista un declive lineal en las habilidades de aprendizaje de una lengua en función de la edad en que se empiece a estudiarla. Esto podría deberse a "the result of developmental factors up to the end of maturation, and of non-developmental factors thereafter". Birdsong también presenta evidencia en contra de la hipótesis del periodo crítico al mencionar la existencia de adultos estudiando una segunda lengua o una lengua extranjera y que han alcanzado una competencia en la misma similar a la de las personas nativas.

El declive gradual a lo largo de toda la vida de las funciones cognitivas puede ser la causa del declive en la habilidad de aprender una lengua. En el capítulo 7 de Birdsong (eds.) (1999) se hace un estudio entre los inmigrantes hispanos y chinos en Estados Unidos, entre 0 y 72 años, que han estado expuestos al idioma al menos durante 10 años, y se demuestra que existe una relación perfectamente lineal entre la edad de llegada al país y el nivel de competencia alcanzado en el inglés como segunda lengua. También se observa que el declive en la competencia permanece constante a lo largo de todas las edades.

Los investigadores han explorado distintas variables que pueden predecir los resultados del aprendizaje de un idioma. Entre estas variables se encuentran la edad de adquisición, el grado de exposición al segundo idioma o al idioma extranjero, la motivación, la integración psicosocial con la cultura del segundo idioma y las estrategias y estilos de aprendizaje.

Las áreas de la lengua más comúnmente investigadas son la pronunciación y la morfosintaxis, y según los resultados de más de dos docenas de estudios experimentales la edad de adquisición es el factor, que de forma más fiable, predice el nivel de competencia que un alumno puede alcanzar. Ver los análisis de Birdsong (2005), y DeKeyser & Larson-Hall (2005).

Johnson y Newport (1989) encontraron una fuerte relación lineal entre la edad de adquisición del idioma inglés, de un grupo de nativos chinos y coreanos, y la corrección con la que hablaban el idioma ($r = - 77$, $p < 0,01$). Este hallazgo fue reproducido por Birdsong y Molis (2001), pero en este caso las personas eran hispanohablantes. La correlación que se obtuvo fue también fuertemente negativa ($r = - 0,77$, $p < 0,0001$). Johnson y Newport (1989) también estudiaron la influencia de la edad en los resultados de la adquisición de un segundo idioma pero esta vez dividieron a los sujetos en dos grupos de ≤ 16 años y >16 años. Este estudio fue repetido por Bialystok y Hakuta (1994) que eligieron la edad de corte a los 20 años y por Birdsong y Molis (2001) que colocaron el punto de corte a varias edades entre 15 y 27,5 años. En todas estas investigaciones se encontraron correlaciones significativas.

En el estudio que Birdsong (2005) realizó teniendo en cuenta la pronunciación y la morfosintaxis de personas que adquirían un segundo idioma llegó a las siguientes conclusiones:

a. En todos los análisis de datos conjuntos de personas que empezaban a adquirir el idioma a distintas edades, el efecto de la edad persistía indefinidamente a lo largo de todas las edades en que se basó la investigación.

b. En los análisis que estudiaban solamente a personas que empezaban a adquirir el idioma tarde, los efectos de la edad sobre los resultados eran significativos.

c. Cuando los estudios se centraban en personas que empezaban a adquirir la lengua a edades tempranas los efectos de la edad eran inconsistentes.

Birdsong hace una exhaustiva recopilación de los numerosos estudios existentes relacionados con la edad y la adquisición de una segunda lengua incluyendo el envejecimiento del cerebro y la distinta forma en que éste procesa el aprendizaje de una primera y una segunda lengua. Resume todos los hallazgos al respecto diciendo que la edad afecta negativamente a la memoria, al procesamiento de los componentes de la lengua y a la producción de la misma. También afirma que este declive es lineal y que se extiende desde que comenzamos nuestra vida de adultos hasta el final de la misma.

> From the cognitive literature, we learn that the associative memory and incremental learning elements of language learning are steadily compromised by age, as are the working memory and processing speed components of language processing and production. It appears that these declines are linear and that they begin in early adulthood and continue throughout the life span.
> (Birdsong, 2006: 34)

El uso de una segunda lengua es menos automática y menos eficiente que el uso de la primera lengua. Por esta razón, las deficiencias de procesamiento aparecen antes y son más pronunciadas en el típico uso de la segunda lengua que en el uso de la primera.

En algunas áreas del cerebro se observa alguna evidencia de la relación entre los cambios morfológicos relacionados con la edad y los procesos cognitivos que tienen lugar en el aprendizaje, la producción y el procesamiento de la segunda lengua.

Tabla 9. 22. Estadísticos de grupo

	Edad	N	Media	Desviación típ.	Error típ. de la media
C-Test. Lexis	<30	90	29,7444	6,48345	,68342
	>=30	61	26,3934	7,44822	,95365
C-Test. Function	<30	90	41,7222	7,42306	,78246
	>=30	61	37,8033	8,95883	1,14706
C-Test	<30	90	71,4667	13,07644	1,37838
	>=30	61	64,1967	15,50357	1,98503
EOI	<30	90	49,6667	8,47708	,89356
	>=30	61	47,2377	9,35486	1,19777

Se quiso analizar si entre los sujetos que habían participado en nuestro estudio y que tenían distintas edades se apreciaba esa diferencia de resultados. Es decir, si la edad era un factor que afectaba a los resultados de los participantes.

Como podemos apreciar en la tabla 9. 22, las medias de los participantes menores de 30 años son más altas que las de los participantes ≥ 30 años. Vemos también que esto es así en C-test de léxico, en el C-test de términos funcionales, en el C-test total y en la prueba global de la EOI, en la que se tienen en cuenta todas las destrezas. Estos resultados son acordes con todos los resultados de las investigaciones que se han hecho al respecto sobre la influencia de la edad en el aprendizaje de una segunda lengua o una lengua extranjera.

La tabla 9.23 nos muestra el estudio que se hizo para saber si la diferencia de medias era o no significativa. Los resultados fueron que existen diferencias significativas entre las medias de los tres modelos de C-test, pero no así en el examen global de la EOI.

Resumiendo, podemos afirmar, basándonos en los datos estadísticos obtenidos, que el C-test fue más fácil para las personas < de 30 años. Sin embargo, aunque los resultados del test de la EOI fueron también más altos

para los sujetos <30 años no se puede afirmar que esa diferencia sea significativa. Esto puede deberse a que los alumnos practican los modelos de examen de la EOI durante todo el año, y puede que la práctica pueda suplir las deficiencias del cerebro, en el procesamiento y la producción de una lengua extranjera, en las personas de mayor edad.

Tabla 9.23. Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bil)	Diferencia de medias	Error típ. de la diferencia	Superior	Inferior
C-Test. Lexis	Se han asumido varianzas iguales	,807	,370	2,933	149	,004	3,35100	1,14238	1,09365	5,60835
	No se han asumido varianzas iguales			2,856	116,702	,005	3,35100	1,17324	1,02739	5,67461
C-Test. Function	Se han asumido varianzas iguales	2,063	,153	2,926	149	,004	3,91894	1,33948	1,27212	6,56577
	No se han asumido varianzas iguales			2,822	112,420	,006	3,91894	1,38852	1,16788	6,67001
C-Test	Se han asumido varianzas iguales	1,599	,208	3,108	149	,002	7,26995	2,33910	2,64786	11,89204
	No se han asumido varianzas iguales			3,008	113,950	,003	7,26995	2,41666	2,48253	12,05736
EOI	Se han asumido varianzas iguales	,975	,325	1,657	149	,100	2,42896	1,46624	-,46835	5,32627
	No se han asumido varianzas iguales			1,625	120,259	,107	2,42896	1,49436	-,52969	5,38762

9.5.12. Relación entre los resultados del C-test, el test de la EOI y el sexo

Ha habido estudios que documentan diferencias individuales en la comprensión lectora de una segunda lengua (ej. Chavez, 2001; Brantmeier, 2001, 2003).

Bügel y Buunk (1996) demostraron que el sexo y el contenido del pasaje eran variables importantes asociadas con diferencias individuales en la comprensión lectora. Brantmeier (2003) indica que la comprensión lectora estaba influenciada de forma significativa por el contenido del pasaje y el sexo de los lectores. Los estudios indicaban que los varones obtenían mejores resultados cuando el contenido de los pasajes era un tema técnico y las mujeres cuando el contenido era un tema relacionado con las humanidades, la cultura o las artes. O´Neil et al. (1993) estudiaron extensamente las diferencias entre sexos por medio de tres modelos del examen de "Graduate Management Admission Test (GMAT)" y llegaron a la conclusión de que los ítems de la comprensión lectora eran mucho más fáciles para los varones que para las mujeres. También descubrieron que los ítems de tipo abstracto eran más fáciles para las mujeres y los de tipo concreto más fáciles para los hombres.

Pae (2004) examinó la influencia del género en un examen de comprensión lectora para alumnos coreanos de inglés como lengua extranjera. Los resultados del estudio indicaron que los ítems que requerían inferencia lógica eran más fáciles para los varones que para las mujeres, mientras que las mujeres obtenían mejores resultados en los ítems con más información contextual. Realizó también un análisis de contenido que reveló que el contenido del pasaje no es un factor fiable para predecir los resultados en un test de comprensión lectora. Por último, en los tests de rellenar huecos los varones consiguieron puntuaciones más altas que las mujeres aunque la investigación se basó en un número limitado de ítems, por lo que estos resultados hay que tomarlos con precaución.

Otros factores que afectan a la actuación de los alumnos en un test son la ansiedad y la confianza en sí mismo, ya que esto último siempre ayuda a conseguir nuestros objetivos. Según Matsuda y Gobel (2003), existen variables estáticas y dinámicas que influyen en la ansiedad. Entre las variables estáticas se encuentran características como la nacionalidad, el género y la primera lengua, que generalmente no cambian con el tiempo. Las variables dinámicas son aquellas que cambian con el tiempo y son diferentes entre los individuos, tales como la autoestima, la motivación o el nivel de competencia de la lengua. Las investigaciones para relacionar la ansiedad con el género han dado distintos resultados. Mejías et al. (1991) encontraron que la ansiedad era mayor entre los varones de habla hispana que entre las mujeres. Sin embargo, estos resultados eran opuestos a los obtenidos por ellos mismos en investigaciones anteriores.

Kitano (2001) encontró que existía correlación entre el género y la ansiedad en los estudiantes japoneses, siendo los varones los que más ansiedad mostraban cuando tenían que hablar en inglés con otros compañeros más competentes. MacIntyre et al. (2002) investigaron como afectaba el sexo, la ansiedad, el nivel de lengua, el deseo de comunicarse y otras variables en los resultados de los tests y encontraron que mientras el nivel de ansiedad de los varones permanecía constante a lo largo de los cursos, el de las mujeres disminuía en el último curso que era cuando aumentaba su deseo de comunicarse. Machida (2001) encontró diferencias significativas entre los sexos, mostrando las mujeres un grado de ansiedad mucho mayor que los varones.

Matsuda y Gobel (2003) estudiaron la posible relación entre la ansiedad creada en la lectura de una lengua extranjera y el género. Los resultados, contrarios a los de Kitano (2001), fueron que no existía una relación significativa entre la ansiedad y el sexo. Sin embargo, el sexo demostró ser una variable significativa para predecir los resultados de los individuos de un curso en las cuatro destrezas de la lengua. El sexo (mujeres) junto con la seguridad en si mismos al hablar en inglés resultaron ser los factores clave del éxito.

Phakiti (2003) examinó si existían diferencias entre el género y los resultados de la comprensión lectora y también si había diferencias en la forma en que los individuos de distinto género usaban las estrategias cognitivas y metacognitivas en el contexto de un test de comprensión lectora. Ambas estrategias se consideran importantes a la hora de comprender un texto en una lengua extranjera.

Las estrategias cognitivas están directamente relacionadas con la lengua objeto de estudio y con el conocimiento que del mundo tengan los individuos, lo cual les permite construir el significado del texto y llevar a cabo la tarea encomendada. Las estrategias cognitivas incluyen hacer predicciones, traducir, resumir, aplicar las reglas gramaticales, relacionar la información del texto con previas experiencias o con conocimiento anterior del tema, y adivinar el significado de palabras o expresiones basándose en el contexto. Las estrategias metacognitivas incluyen tareas de evaluación, planificación y comprobación de las tareas para decidir lo que se debe hacer y cuando y como debe hacerse.

Los hallazgos sobre la relación que existe entre el género de los participantes en un test de comprensión lectora y sus resultados no son consistentes. Por una parte, tenemos estudios como los de Wen y Johnson (1997) que encontraron que las mujeres obtenían mejores resultados que los hombres en un test de competencia nacional estandarizado. Chavez (2001) encontró que independientemente de los temas y el género del texto, las mujeres obtenían puntuaciones más altas que los varones en un test de elección múltiple de comprensión lectora.

Por otra parte, tenemos otros estudios que demuestran que los varones obtienen mejores resultados que las mujeres. (Ej. Boyle, 1987). Scarcella y Zimmerman (1998) encontraron que los varones obtenían notas más altas que las mujeres en diferentes tests de vocabulario. Bügel y Buunk (1996) hallaron que cuando los textos eran neutrales en cuanto a género y contenido, los

varones sacaban mejores puntuaciones que las mujeres en los tests de comprensión lectora.

En cuanto al uso de diferentes estrategias en la comprensión lectora, Young y Oxford (1997) no encontraron, en general, diferencias significativas en el uso de estrategias entre varones y mujeres. Sin embargo, los varones monitorizaban el ritmo y las estrategias de lectura, mientras que las mujeres se centraban en resolver los problemas de vocabulario. Estos datos les hizo sugerir que algunas estrategias podrían estar relacionadas con el género.

Los resultados de la investigación de Phakiti (2003) señalan que no existen diferencias entre los varones y las mujeres en las puntuaciones obtenidas en la comprensión lectora ni tampoco en el uso de estrategias cognitivas. Sin embargo, se encontró que los varones usaron las estrategias metacognitivas de forma mucho más frecuente que las mujeres. Este estudio proporciona evidencia empírica que aunque la diferencia de sexos no juega un papel significativo en los resultados de la comprensión lectora, sin embargo, puede jugar un papel importante en el uso de estrategias cognitivas y metacognitivas. Así mismo se halló que los individuos que obtenían los mejores resultados hacían más uso de las estrategias cognitivas y metacognitivas que los que obtenían peores resultados.

Tabla 9.24. Estadísticos de grupo

	Sexo	N	Media	Desviación típ.	Error típ. de la media
C-Test. Lexis	Varón	39	30,3077	6,74402	1,07991
	Mujer	112	27,7232	7,07516	,66854
C-Test. Function	Varón	39	42,8974	6,61257	1,05886
	Mujer	112	39,1786	8,60360	,81296
C-Test	Varón	39	73,2051	12,49037	2,00006
	Mujer	112	66,9018	14,85180	1,40336
EOI	Varón	39	50,7564	8,44691	1,35259
	Mujer	112	47,9643	8,96646	,84725

También se ha querido analizar en nuestro estudio si existían diferencias entre los resultados obtenidos por ambos sexos tanto en el C-test como en el test de la EOI.

La tabla 9.24 nos muestra que tanto en todos los modelos de C-test como en el test global de la EOI las medias obtenidas por los varones son mayores que las obtenidas por las mujeres. Este resultado puede parecer, a primera vista, sorprendente, ya que en la mayoría de la literatura basada en el género, las mujeres parecen tener más facilidad para aprender una segunda lengua que los varones (ver Chavez, 2001).

Sin embargo, teniendo en cuenta por una parte, la investigación de O´Malley y Chamot (1990) que hallaron que los estudiantes que obtienen mejores notas son los que más utilizan las estrategias metacognitivas, y por otra parte la investigación de Phakiti (2003) que halló un mayor uso de estrategias metacognitivas por los varones que por las mujeres, podemos llegar a la conclusión de que ésta puede ser una de las causa de que en nuestro estudio los resultados de los varones sean mejores que los de las mujeres. Estos resultados apoyarían el modelo de habilidad lingüística de Bachman y Palmer (1996) que dice que las estrategias cognitivas y metacognitivas ayudan a explicar la actuación de los alumnos cuando obtienen buenos o malos resultados.

Además, tenemos que tener en cuenta el factor de la ansiedad que, como ya hemos visto, también es una variable que influye negativamente en las puntuaciones finales de los que hacen un test.

Merece la pena resaltar que la mayoría de las investigaciones psicológicas han llegado a la conclusión de que las mujeres sufren mayor ansiedad en los exámenes que los varones (ver Zeidner, 1998). Algunos estudios empíricos como los de Couch et al., (1983) apoyan la hipótesis de que las mujeres interpretan y responden a los exámenes de diferente manera que los varones. Los varones pueden percibir una situación de examen como un reto personal. Las mujeres, por el contrario, la pueden percibir como una

amenaza originándoles estados de miedo, preocupación y baja autoestima, lo cual puede afectar de forma negativa a sus actuaciones en los exámenes.

Al estudiar si las diferencias de medias entre los géneros eran o no significativas (tabla 9.25), encontramos que existen diferencias significativas en los tres modelos de C-test (C-test de léxico, C-test de términos funcionales y C-test global), ya que la $t_{crit} = 1,96$ para un $p< 0,05$, con lo que en este caso $t_{observada} > t_{crit}$ con $p< 0,05$ y $gl = 149$. Sin embargo, vemos que esas diferencias de medias entre los sexos, que muestra la tabla 9.24, en el test global de la EOI no son significativas puesto que en este caso la $t_{observada} < t_{crit}$.

Estos resultados son similares a los hallados al estudiar la influencia que la edad tenía en los resultados de los tests. En ese caso también existían diferencias significativas en las medias de los C-tests pero no en las medias del EOI. Como ya explicamos entonces, esto puede deberse a la práctica que los alumnos realizan a lo largo del curso con los modelos de tests que se van a encontrar en el examen final. Ello puede haber originado que los alumnos hayan utilizado diferentes estrategias en el test de la EOI que en el C-test, puesto que la batería de tests de la EOI era conocida por los alumnos, mientras que el C-test era una técnica de examen desconocida por ellos. La práctica de los modelos de tests de la EOI puede también disminuir la ansiedad que los alumnos sufren en un examen. Por el contrario, al hacer el C-test los alumnos se encontraron con un test que no conocían, lo cual puede haber creado un mayor estado de ansiedad entre los participantes en el estudio, afectando así a las notas finales.

Tabla 9. 25. Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bil)	Diferencia de medias	Error típ. de la diferencia	Superior	Inferior
C-Test. Lexis	Se han asumido varianzas iguales	,003	,954	1,988	149	,049	2,58448	1,30005	,01556	5,15340
	No se han asumido varianzas iguales			2,035	69,227	,046	2,58448	1,27010	,05085	5,11810
C-Test. Function	Se han asumido varianzas iguales	1,887	,172	2,457	149	,015	3,71886	1,51387	,72743	6,71029
	No se han asumido varianzas iguales			2,786	85,798	,007	3,71886	1,33495	1,06498	6,37275
C-Test	Se han asumido varianzas iguales	,772	,381	2,373	149	,019	6,30334	2,65631	1,05444	11,55224
	No se han asumido varianzas iguales			2,580	78,143	,012	6,30334	2,44329	1,43926	11,16742
EOI	Se han asumido varianzas iguales	,000	,994	1,699	149	,091	2,79212	1,64303	-,45452	6,03877
	No se han asumido varianzas iguales			1,749	69,982	,085	2,79212	1,59604	-,39108	5,97533

9.6. Conclusiones

Resumiendo, podemos afirmar que los alumnos de la EOI son alumnos motivados para aprender el idioma inglés, ya que la mayoría de ellos afirman usar el idioma inglés fuera de clase. Muchos aseguran que leen, hablan o ven películas en inglés.

Hemos visto que existe una correlación entre los que usan el idioma o leen fuera de clase y los resultados del C-test, obteniendo mejores notas los que practican el idioma fuera de clase. También se ha demostrado que existen diferencias significativas entre las puntuaciones del C-test de los alumnos < 30 años y los ≥ 30 años, alcanzando notas más altas las personas < 30 años. Por otra parte, el sexo también parece influir en los resultados del C-test y hemos visto que son los varones los que alcanzan las medias más altas. Así mismo, se ha demostrando empíricamente que las diferencias de medias entre los varones y las mujeres son significativas.

Sin embargo, ni la edad ni el género parecen influir de forma significativa en los resultados del test global de la EOI.

Capítulo 10

DISCUSIÓN DE LOS RESULTADOS Y CONCLUSIONES

10.1. Discusión de los resultados

Teniendo en cuenta que las dos versiones del C-test se crearon usando cuatro textos nuevos y que no se realizó ningún análisis de los términos de antemano, los coeficientes de fiabilidad son satisfactorios. Además, los alumnos representaban una muestra en la que los conocimientos de la lengua se movían en un rango muy estrecho al pertenecer todos a un mismo nivel de la EOI y esto, de acuerdo con Weir (2005: 32), "effectively limits estimates of internal consistency level". Por otra parte los textos resultaron ser demasiado fáciles para la mayoría de los candidatos, lo que limita también el coeficiente de fiabilidad.

Los datos de esta investigación nos indican que la dificultad de los cuatro textos variaba debido principalmente al léxico que contenía cada pasaje. El C-test de léxico es más difícil que el C-test funcional, aunque la contribución del léxico a la medida de la competencia general de la lengua es también mayor, como lo demuestra que el coeficiente de correlación del C-test de léxico con el test global de la EOI es el más alto. Por otra parte, como afirma Bachman (1990: 37), es muy difícil y probablemente no deseable "to develop tests in which all he tasks or items are at the exact level of difficulty appropriate for the individuals being tested".

Se ha demostrado que existen diferencias significativas entre el C-test de términos léxicos y el C-test de términos funcionales, tanto a nivel de C-test global como a nivel de textos o super-ítems. Esto se puede utilizar para determinar la dificultad o nivel de lengua de un pasaje.

Los resultados en este estudio indican que hay diferencias significativas entre los super-ítems de los C-tests A y B. Puesto que los textos son los mismos y los dos grupos de estudiantes son homogéneos, la única variable que puede originar esas diferencias es el punto donde se empiece a dañar las

palabras. Podemos inferir por lo tanto que el punto en el que se empiece a mutilar un texto afecta a la media de las puntuaciones de los super-ítems. Sin embargo, estas diferencias entre las medias de cada texto se neutralizan y desaparecen cuando se suman las puntuaciones de todos los super-ítems que forman el test. Se demostró que no existen diferencias significativas entre las medias de los dos modelos de C-test (C-test A y C-test B). Puede decirse que el C-test A y el C-test B son tests equivalentes o dos formas paralelas del mismo test. El procedimiento de mutilación de palabras parece afectar de forma diferente a las palabras individuales de un texto y a los subtests, pero como el C-test consta de varios textos el efecto se anula.

Este hallazgo es muy importante, puesto que la elaboración del C-test se simplifica muchísimo al no tener que tener en cuenta en qué punto se empieza a mutilar un texto, ya que empecemos donde empecemos no vamos a tener diferencias significativas en los resultados del test. Por otra parte, el hallazgo también demuestra la importancia de construir el C-test con varios textos distintos, y no solamente para paliar el efecto que el contenido o el tema pueda tener sobre los candidatos sino para evitar o neutralizar las diferencias que pueda haber en los resultados de los C-tests que se pueden crear dependiendo de donde empecemos a mutilar los textos.

Con el propósito de determinar como los dos modelos del C-test correlacionan con un criterio exterior, las puntuaciones de los C-tests se correlacionaron con las de los tests de la EOI.

Con respecto a la validez de los dos C-tests, los valores tan significativos y satisfactorios de los coeficientes de correlación obtenidos entre los C-tests y el test global de la EOI, parecen indicar un alto grado de asociación entre ellos, sugiriendo que los dos tests miden el mismo constructo, es decir, la competencia general de la lengua. Los coeficientes de correlación entre el C-test y la batería de tests que forman el examen de comprensión lectora son también significativos pero más bajos. Como ya se ha dicho, esto puede deberse al hecho de que diferentes técnicas parecen medir diferentes aspectos de la comprensión lectora (Kobayashi, 2002: 211; Bensoussan, 1984: 56; y

Weir, 2005: 32). Sin embargo, la inaceptablemente baja correlación entre el test de casar extractos con títulos y el test de elección múltiple debe tenerse en cuenta y tratar de encontrar las razones por lo que esto ocurre, así como analizar de qué forma contribuyen a la habilidad total de la comprensión lectora.

La correlación del C-test con el test global de la EOI es mayor que incluso la de otros tests que forman el test global de la EOI, por ejemplo, los tests de elección múltiple o el test de casar títulos con párrafos. Ello puede deberse al hecho de que el C-test mide la competencia general de la lengua mientras que otros tests miden rasgos más específicos. La correlación del *cloze* con el test de la EOI es mayor que la del C-test, pero esto es lógico, ya que el *cloze* es parte integrante de la batería de tests de la EOI.

Por lo que respecta a la recuperación de las palabras ésta dependerá de factores tales como el nivel de competencia del alumno, la dificultad del texto, la existencia de flexiones o sufijos añadidos en la formación de esos términos, si las palabras son pautadas o no, de la longitud de los términos, del primer idioma del candidato, y sobre todo de la frecuencia de las palabras tanto en el vocabulario activo como en el pasivo del examinando. También se ha observado que la edad y el género influyen en los resultados del C-test demostrando empíricamente que son los varones y los menores de 30 años los que obtienen mejores puntuaciones con diferencias de medias significativas en ambos casos.

Basándonos en los datos obtenidos en el cuestionario podemos afirmar que existe una correlación positiva entre los alumnos que usan o leen en inglés fuera de clase y los resultados del C-test.

En cuanto a la validez aparente del C-test, éste no parece ser un test que origine fuertes rechazos ni grandes adhesiones, de acuerdo con las opiniones de los alumnos de la EOI recogidas en el cuestionario. Esto puede deberse a la falta de experiencia con el procedimiento y el método de examen por parte de los alumnos que realizaron los tests.

10.2. Conclusiones

En términos generales, los resultados de este estudio indican que:

1. Existe una correlación muy significativa entre el C-test y el c*loze* además de entre el C-test y la batería de tests usados en la EOI.

2. No existen diferencias significativas entre los dos modelos de C-test creados, es decir, que ambos tests son equivalentes.

3. Existen diferencias significativas entre los dos modelos de super-ítems dependiendo del punto donde se empiece a mutilar las palabras de los textos, es decir, dependiendo de si se empieza a mutilar los textos en la segunda o en la tercera palabra después del primer punto de cada texto.

4. Existen diferencias significativas entre las palabras léxicas y las funcionales recuperadas tanto a nivel de texto o super-ítem como a nivel de C-test.

5. Las palabras pautadas se recuperan más fácilmente que las no pautadas.

6. Las correlaciones entre el C-test y el test de competencia global de la EOI son muy notables.

El hecho de que los coeficientes de correlación entre el C-test y los tests de la EOI sean significativos, los coeficientes de fiabilidad obtenidos en el estudio sean respetables y la consistencia interna entre los subtests que forman el C-test sea relativamente alta, sugiere que el C-test podría sustituir al

cloze en la evaluación de la habilidad de la comprensión lectora. También podría ser una buena opción el incluirlo en la batería de técnicas usadas en la EOI, ya que ayudaría a aumentar la fiabilidad y mejorar la calidad de la batería de tests en su conjunto.

Resumiendo, podemos afirmar, con los resultados que se han obtenido en esta investigación, que el C-test es una buena alternativa al *cloze*, ya que ha demostrado ser una técnica económica, objetiva, fiable, válida y más fácil de elaborar y de corregir que el *cloze*.

10.3. Planes para futuras investigaciones

Esta investigación se ha realizado con alumnos de nivel intermedio. Es posible que los resultados no puedan generalizarse para los alumnos cuyo nivel de inglés sea elemental o muy avanzado. Una investigación que tendría gran utilidad para la EOI es el estudiar hasta qué punto el C-test se puede utilizar en exámenes a gran escala y para todos los niveles de competencia.

El C-test podría ser una solución al problema de los exámenes de clasificación, ya que es un test fácil de preparar y corregir, y parece ser un método válido para medir la competencia global de una lengua. Ya hemos visto que presenta una alta correlación con el test de la EOI. Tendríamos que demostrar si el C-test es capaz de discriminar entre grupos de distinto nivel. Si eso fuera así, este modelo de examen podría jugar un papel importante a la hora de determinar a que curso debe acceder un aspirante a ser alumno oficial de las Escuelas Oficiales de Idiomas, y de esta forma facilitar los siempre complejos y a veces subjetivos exámenes anuales de clasificación. Estas pruebas se acortarían considerablemente y el proceso de elaboración, administración y corrección sería más económico tanto a nivel de esfuerzo como de tiempo empleado.

BIBLIOGRAFÍA

Aborn, M., Rubenstein, M. y Sterling, T. D. (1959). Sources of contextual constraint upon words in sentences. *Journal of Experimental Psychology* 57: 171-180.

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing* 24 (1) 7-36.

Aitchison, J. (1989). *Words in the mind*. Oxford: Basil Blackwell.

Alcaraz, E. y Ramón, J. (1980). *La evaluación del inglés. Teoría y práctica*. Madrid: SGEL, S. A.

Alderson, J. C. (1979a). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly* 13 (2): 219-227.

Alderson, J. C. (1979b). The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading* 2 (2): 108-119.

Alderson, J. C. (1980). Native and non-native speaker performance on cloze tests. Language Learning 30 (1): 59-76.

Alderson, J. C. (1984). Reading in a foreign language: a reading problem or a language problem? En J. C. Alderson y A. H. Urquhart (eds.), *Reading in a foreign Language* 1-24. London: Longman

Alderson, J. C. (1990). Testing Reading Comprehension Skills (Part I). *Reading in a foreign Language* 6 (2): 425-438.

Alderson, J. C. (1991a). Bands and scores. En J. C. Alderson y B. North. *Language Testing in the 1990s: The Communicative Legacy* 71-86. London: Mcmillan.

Alderson, J. C. (1991b). Dis-sporting Life. Response to Alistair Pollit´s paper. En J. C. Alderson y B. North. (1991). *Language Testing in the 1990s: The Communicative Legacy* 60-67. London: Mcmillan.

Alderson, J. C. (2000a). *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2000b). Technology in testing: the present and the future. *System* 28: 593-603.

Alderson, J. C. (2002). Testing proficiency and achievement: principles and practice. En James A. Coleman, Rüdiger Grotjan & Ulrich Raatz (eds.), *University language testing and the C-test* 15-30. Bochum: AKS-Verlag.

Alderson, J. C. (2007a). Judging the Frequency of English Words. *Applied Linguistics* 28 (3): 383- 409.

Alderson, J. C. (2007b). The CEFR and the Need for More Research. *The Modern Language Journal* 91: 659-663.

Alderson, J. C. (eds.). (1985). Lancaster practical Papers in English Language Education. Vol.6. *Evaluation*. Oxford: Pergamon Press.

Alderson, J. C. et al. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference. The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly* 3 (1): 3-30.

Alderson, J. C. y Beretta, A. (1992). *Evaluating Second Language Education*. Cambridge: Cambridge University Press.

Alderson, J. C. y Hamp-Lyons, L. (1996). TOEFL preparation course: a study of Washback. *Language Testing* 13 (3): 280-297.

Alderson, J. C. y Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing* 22 (3): 301-320.

Alderson, J. C. y North, B. (1991) *Language Testing in the 1990s: The Communicative Legacy*. London: Mcmillan.

Alderson, J. C. y Urquhart, A. H. (1983). The effect of student background discipline on comprehension: a pilot study. En D. Porter y A. Hughes (eds.), *Current Developments in Language Testing*: 121-128 London: Academic Press.

Alderson, J. C. y Urquhart, A. H. (1984). ESP tests: the problem of student background discipline. En T. Culhane et al. (eds.), *Practice and problems in language testing*. Occasional Papers 29, 1-13. Colchester: University of Essex.

Alderson, J. C. y Urquhart, A. H. (1985). The effect of students´ academic discipline on their performance on ESP reading tests. *Language testing* 2 (2): 192-204

Alderson, J. C. y Wall, D. (1993). Does washback exist?. *Applied Linguistics* 14 (2): 115-129.

Alderson, J. C. y Windeatt, S. (1991). Computers and Innovation in Language Testing. En J. C. Alderson & B. North. *Language Testing in the 1990s: The Communicative Legacy* 226-236. London: Mcmillan.

Alderson, J. C., Clapham, C. y Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson. J. C. y Banerjee, J. (2002). Language testing and assessment. *Language teaching* 35: 79-113.

Allwright, J. y Allwright , R. (1977). An approach to the teaching of medical English. En S. Holden (eds.), *English for Specific Purposes* 58-62. Oxford: Modern English Publications.

Allwright, R. (1982). Perceiving and pursuing learner's needs. En M. Geddes y G. Sturtridge (eds). *Individualisation* 24-31. Oxford: Modern English Publications.

Álvarez Méndez, J. M. (2001). *Evaluar para conocer, examinar para excluir*. Madrid: Ediciones Morata.

Álvarez Méndez, J. M. (2003). *La evaluación a examen. Ensayos críticos*. Madrid: Miño y Dávila (eds.).

Amengual Pizarro, M. (2003). A Study of Different Composition Elements that Raters Respond To. *Estudios Ingleses de la Universidad Complutense de Madrid*, 11: 53-72.

Amengual Pizarro, M. (2004). Reliability Concerns on Evaluating ESL compositions, in: Carretero Lapeyre et al. (Eds) *Estudios de Lingüística aplicada a la comunicación*. Madrid: CERSA, 15-27.

Amengual Pizarro, M. (2005). Posibles sesgos en el examen de Selectividad, En: Herrera Soler & García Laborda (Eds.) *Estudios y Criterios para una evaluación de calidad.* (Valencia, Universidad Politécnica de Valencia), 121-148.

Amengual Pizarro, M. (2006). Análisis de la prueba de inglés de Selectividad de la Universitat de les Illes Balears. *Revista Ibérica*, Asociación Europea de Lenguas para Fines Específicos AELFE 11: 29-59.

Amengual Pizarro, M. (2009). Does the English Test in the Spanish University Entrance Examination Influence the Teaching of English?. *English Studies*. 90 (5) Routledge: Oxfordshire UK, 582- 598.

Anderson, J. N. (1991). Individual differences in strategy use in second language reading and testing. *Modern Language Journal* 75: 460-472.

Arnaud, P. J. L. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. En T. Culhane et al. (eds.), *Practice and problems in language testing*. Occasional papers 29: 14-28. Colchester: University of Essex.

Arnaud, P. J. L. y Béjoint, H. (eds.), (1992). *Vocabulary and Applied Linguistics*. London: Macmillan.

Babaii, E. y Ansary, H. (2001). The C-test: a valid operationalization of reduced redundancy principle?. *System* 29 (2): 209-219.

Babaii, E. y Moghaddam, M. J. (2006). On the interplay between test task difficulty and macro-level processing in the C-test. *System* 34: 586-600.

Bachman, L. F. (1985). Performance on the cloze test with fixed-ratio and rational deletions. *TESOL Quarterly* 19 (3): 535-556.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. (1998). Language testing – SLA research interfaces, En L. F. Bachman y A. D. Cohen. *Interfaces Between Second Language Acquisition and Language Testing Research*, 177-195. Cambridge: Cambridge University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing* 17 (1): 1-42.

Bachman, L. F. (2004). *Statistical Analysis for Language Assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. y Cohen, A. D.(1998). Language testing – SLA interfaces: An update. En L. F. Bachman y A. D. Cohen. *Interfaces Between Second Language Acquisition and Language Testing Research*, 1-31. Cambridge: Cambridge University Press.

Bachman, L. F. y Palmer, A. S. (1996). *Language Testing in Practice*: *Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Barcroft, J. (2007). Effects of Opportunities for Word Retrieval During Second Language Vocabulary Learning. Language Learning 57 (1): 35-56.

Barnett, M. A. (1986). Syntactic and lexical/semantic skills in foreign language reading: importance and interaction. *Modern Language Journal* 70: 343 - 349

Bedford, J. (2002). Washback – the Effect of Assessment on ESOL Teaching and Learning. (Documento disponible en Internet en: http://www.tki.org.nz/r/esol/esolonline/teachers/prof_read/jenni_bedford/home_e.php).

Beglar, D. y Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing* 16 (2): 131-162.

Bernhardt, E. B. y Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*. 16 (1): 15-34

Bensoussan, M. y Ramraz, R. (1984). The fill-in test: a modified multiple-choice cloze technique to test Reading Comprehension of English as a Foreign Language. En T. Culhane et al. (eds.), Practice and problems in language testing. *Occasional papers* 29, 44-65. Colchester: University of Essex.

Bialystok, E. y Hakuta, K. (1994). *In other words: the science and psychology of second language acquisition*. New York: Basic books.

Birdsong, D. (1999). Introduction: Whys and why nots of the Critical period hypothesis. En D. Birdsong (eds.), *Second Language Acquisition and the Critical Period Hypothesis*. Mahwah, NJ: Erlbaum.

Birdsong, D. (2005). Interpreting age effects in second language acquisition. En J. F. Kroll y A.M. B. de Groot (eds.). *Handbook of bilingualism: Psycholinguistic approaches*, 109-127. New York: Oxford University Press.

Birdsong, D. (2006). Age and Second Language Acquisition and Processing. A Selective Overview. *Language Learning* 56 (1): 9-49.

Birdsong, D. (eds.), (1999). *Second Language Acquisition and the Critical Period Hypothesis*. Mahwah, NJ: Erlbaum.

Birdsong, D. y Molis, M. (2001). On the evidence of maturational effects in second language acquisition. *Journal of memory and language* 44: 235-249.

Biskup, D. (1992). L1 Influence on Learners´ Renderings of English Collocations: A Polish/German Empirical Study. En Arnaud, J. L. y Béjoint, H. (eds.), (1992). *Vocabulary and Applied Linguistics*, 85-94. London: Mcmillan

Bogaards, P. (2000). Testing L2 Vocabulary Knowledge at a High Level: the Case of the *Euralex French Tests*. *Applied Linguistics* 21 (4): 490-516.

Boyle, J. P. (1987). Sex differences in listening vocabulary. *Language Learning* 37: 273-284.

Bradshaw, J. (1990). Test-takers´ reactions to a placement test. *Language Testing* 7 (1): 13-30.

Braine, G. (2001). When an exit test fails. *System* 29: 221-234.

Brantmeier, C. (2001). Second language reading research on passage content and gender: Challenges for the intermediate-level curriculum. *Foreign Language Annals* 34 (4): 325-333.

Brantmeier, C. (2003). Beyond linguistic knowledge: Individual differences in second language reading. *Foreign Language Annals* 36 (1): 33-43.

Brindley, G. (1998). Describing language development?. Rating scales and SLA. En L. F. Bachman y A. D. Cohen. *Interfaces Between Second Language Acquisition and Language Testing Research*, 112-140. Cambridge: Cambridge University Press

Brindley, G. (2001). Outcomes-based assessment in practice: some examples and emerging insights. *Language Testing* 18 (4): 393-407.

Broadfoot, P. M. (1996). *Education, Assessment and Society. A Sociological Analysis*. Buckingham: Open University Press.

Broadfoot, P. M. (2005). Dark alleys and blind bends: Testing the language of learning. *Language Testing* 22 (2): 123-141.

Brown, J. D. (1993a). *Understanding Research in Second Language Learning*. Cambridge: Cambridge University Press.

Brown, J. D. (1993b). What are the characteristics of natural cloze tests?. *Language Testing* 10 (3): 93-116.

Brown, J. D. (2001). *Using surveys in language programmes*. Cambridge, UK: Cambridge University Press.

Brown, J. D. y Rogers, T. S. (2002). *Doing Second Language Research*. Oxford: Oxford University Press.

Buck, G. (1992). Translation as a language testing process: Does it work?. *Language Testing* 9 (2): 123-148.

Bügel, K. y Buunk, B. P. (1996). Sex differences in foreign language text comprehension: The role of interests and previous knowledge. *Modern Language Journal* 80 (1): 15-31.

Butler, C. (1985). *Statistics in Linguistics*. Oxford y New York: Basil Blackwell.

Carnine, D., Kameenui, E. J., y Goyle, G. (1984). Utilization of contextual information in determining the meaning of unfamiliar words. *Reading Research Quarterly* 19 (2): 188-204.

Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing* 23 (3): 269-289.

Carrell, P. (1987a). Readability in ESL. *Reading in a Foreign Language* 4: 21-40.

Carrell, P. (1987b). Content and formal schemata in ESL reading. *TESOL Quarterly* 21: 461-481.

Carroll, B. J. (1991). Response to Don Porter's Paper: "affective Factors in Language testing". En J. C. Alderson y B. North. *Language Testing in the 1990s: The Communicative Legacy* 41-45. London: Mcmillan.

Carter, R y MacCarthy, M. (eds.). (1988). *Vocabulary and Language teaching*. London: Longman.

Carter, R. (1988). Vocabulary, cloze and discourse: an applied linguistic view. En R. Carter Y M. McCarthy (eds.), *Vocabulary and language teaching*, 161-180. London: Longman.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing* 20 (4): 369-383.

Chalhoub-Deville, M. y Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. System 28: 523-539.

Channel, J. (1988). Psycholinguistic considerations in the study of L2 vocabulary acquisition. En R. Carter Y M. McCarthy (eds.), *Vocabulary and language teaching*, 83-96. London: Longman.

Chapelle, C. A, y Douglas, D. (2006). Assessing language through computer technology. Cambridge: Cambridge University Press.

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research?. *Second Language research* 10: 157-187.

Chapelle, C. A. (1998). Construct definition and validity enquiry in SLA research. En L. F. Bachman y A. D. Cohen. *Interfaces Between Second Language Acquisition and Language Testing Research*, 32-70. Cambridge: Cambridge University Press.

Chapelle, C. A. (2006). DIALANG: A diagnostic language test in 14 European languages. Test Review. Language Testing 23 (4): 544-550.

Chapelle, C. A. y Abraham, R. (1990). Cloze method: what difference does it make?. *Language Testing* 7: 121-146.

Chavez, M. (2001). *Gender in the language classroom*. Boston: McGraw Hill.

Chen, L. (2004). On text structure, language proficiency, and reading comprehension test format interactions: a reply to Kobayashy, 2002. *Language Testing* 21 (2): 228-234.

Cheng, L., Rogers, T. y Hu, H. (2004). ESL/EFL instructors´ classroom assessment practices: purposes, methods, and procedures. *Language Testing* 21 (3): 360-389.

Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing* 25 (1): 39-62.

Clapham, C. (2000). Assessment for academic purposes: where next?. *System* 28: 511-521.

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. En Jones y Spolsky, 10-24.

Clark, J. L. D. (1978a*). Direct Testing of Speaking Proficiency: Theory and Application*. Princeton, NJ: Educational Testing Service.

Clark, J. L. D. (1978b). Interview testing research at Educational Testing Service. En Clark 1978a, 211-228.

Clark, J. L. D. (1983). Language Testing: Past and Current Status - Directions for the Future. *The Modern Language Journal* 67: 421-443.

Clarke, M. (1979). Reading in English and Spanish: evidence from adult ESL students. *Language Learning* 29: 121-150.

Cleary, C. (1988). The C-Test in English: left-hand deletions. *RELC Journal* 19 (2): 26-35.

Coady, J. y Huckin, T. (eds). (1997). *Second Language Vocabulary Acquisition*. Cambridge: Cambridge University Press.

Cohen, A. D. (1998). Strategies and processes in test taking and SLA. En L. F. Bachman y A. D. Cohen. (1998). *Interfaces Between Second Language Acquisition and Language Testing Research*, 90-111. Cambridge: Cambridge University Press.

Cohen, A. D. y Upton, T. A. (2007). "I want to go back to the text": Response strategies on the reading subtest of the new TOEFL®. *Language Testing* 24 (2): 209-250.

Coleman, J. A. (2002). The European Proficiency Survey: an overview of findings. En James A. Coleman, Rüdiger Grotjan & Ulrich Raatz (eds.), *University language testing and the C-test,* 201-206. Bochum: AKS-Verlag.

Connelly, M. (1997). Using C-Test in English with Postgraduate Students. *English for Specific Purposes* 16 (2): 139-150.

Consejo de Europa (2001). Marco Común Europeo de Referencia para las Lenguas: aprendizaje, enseñanza, evaluación. Madrid: Instituto Cervantes.

Corrigan, R. (2007). An Experimental Analysis of the Affective Dimensions of Deep Vocabulary Knowledge Used in Inferring the Meaning of Words in Context. *Applied Linguistics* 28 (2): 211-240.

Couch, J., Garber, T. B. y Turner, W. E. (1983). Facilitating and debilitating test anxiety and academic achievement. *Psychological Reports* 33: 237-244.

Culhane, T., Klein-Braley, C. y Stevenson, D. K. (eds.), (1982). *Practice and problems in language testing*. Colchester: University of Essex.

Culhane, T., Klein-Braley, C. y Stevenson, D. K. (eds.), (1984). *Practice and problems in language testing*. *Occasional papers*. Colchester: University of Essex.

Cumming, A. y Berwick, R. (eds.) (1996). *Validation in Language Testing*. Clevedon: Multilingual Matters Ltd.

Daller, H. y Phelan, D. (2006). The C-Test and TOEIC® as measures of students´ progress in intensive short courses in EFL. En R. Grotjahn, *The C-Test: Theory, Empirical Research, Applications,* 101-120. Frankfurt: Peter Lang.

Dastjerdi, H. V. y Talebinezhad, M. R. (2006). Chain-preserving deletion procedure in cloze: a discoursal perspective. *Language Testing* 23 (1): 58-72.

David, G. (2007). Investigating the performance of alternative types of grammar items. *Language Testing* 24 (1): 65-87.

Davidson, F. y Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language teaching* 40: 231-241.

Davies, A. (1975). Two tests of speed reading. En R. L. Jones y B. Spolsky (eds.). (1975). *Testing Language Proficiency*. Arlinton, Va.: Center for Applied Linguistics.

Davies, A. (1991). Language Testing in the 1990s. En J. C. Alderson y B. North, *Language Testing in the 1990s: The Communicative Legacy,* 136-150. London: Mcmillan.

Davies, A. (1997). Demands of being professional in language testing. Language Testing 14 (3): 328-339.

Davies, A. (2003). Three heresies of language testing research. *Language Testing* 20 (4): 355-368.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. y Mcnamara, T. (1999) *Dictionary of Language Testing*. Studies in Language Testing, Vol. 7. Cambridge: UCLES/Cambridge University Press.

Decy. E. L. y Ryan, R. M. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist* 55 (1): 68-78.

DeKeyser, R. y Larson-Hall, J. (2005). What does the critical period really mean?. En J. F. Kroll y A. M. B. de Groot (eds.). *Handbook of*

bilingualism: Psycholinguistic approaches, 109-127. new York: Oxford University Press.

Deville, C. y Chalhoub-Deville, M. (1993). Modified scoring, traditional item analysis and sato´s caution index used to investigate the reading recall protocol. *Language Testing* 10 (2): 117-132.

Dörnyei, Z. (2003). *Questionnaires in Second Language Research: Construction, Administration and Processing*. London: Lawrence Erlbaum Associates.

Dörnyei, Z. y Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing* 9 (2): 187-206.

Douglas, D. (1998). Testing methods in second language research. En L. F. Bachman y A. D. Cohen. (1998). En *Interfaces Between Second Language Acquisition and Language Testing Research*, 141-155. Cambridge: Cambridge University Press.

Douglas, D. (2000). *Assessing Languages for specific Purposes*. Cambridge: Cambridge University Press.

Eckes, T. y Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing* 23 (3): 290-325.

Eckes, T., Ellis, M., Kalberzina, V., Pzorn, K., Spinger, C., Szollás, K. y Tsagari, C. (2005). Progress and problems in reforming public language examinations in Europe: cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. *Language Testing* 22 (3): 355-377.

Elder, C. et al. (2007). Evaluating rater responses to an on line training program for L2 writing assessment. Language Testing 24 (1): 37-64.

Ellis, N. C. (1997). Vocabulary acquisition: word structure, collocation, word-class, and meaning. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, 122-139. Cambridge: Cambridge University Press.

Ellis, N. C. Y Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning* 43: 559-617.

Esteban García, M. (2005). El C-Test y la prueba de selectividad e inglés, En: Herrera Soler & García Laborda (eds.). *Estudios y Criterios para una*

evaluación de calidad. (Valencia, Universidad Politécnica de Valencia), 165-185.

Esteban García , M. (2008). Las pruebas de cierre en la enseñanzade lenguas extranjeras. *Educación y Futuro* 19; 73-90. Madrid: CES Don Bosco.

Falk, B. (1984). Can grammatical correctness and communication be tested simultaneously?. En T. Culhane et al. (eds.), *Practice and problems in language testing*. Occasional papers 29: 44-65. Colchester: University of Essex.

Figueras, N. (2007). The CEFR, a Lever for the Improvement of Language Professionals in Europe. *The Modern Language Journal* 91: 673-675.

Figueras, N., North, B., Takala, S., Verhelst, N. y Van Avermaet, P. (2005). Relating examinations to the Common European Framework: a manual. *Language Testing* 22 (3): 261-279.

Fotos, S. (1991). The Cloze Test as an Integrative Measure of ELF Proficiency: A Substitute for Essays on College Entrance Examination?. *Language Learning* 41 (3): 313-336.

Fox, J. (2004). Test decisions over time: tracking validity. *Language testing* 21 (4): 437-465.

Frasson, A. (1984). Cramming or understanding?. Effects of intrinsic and extrinsic motivation on approach to learning and test performance. En J. C. Alderson y A. H. Urquart (eds.), *Reading in a foreign language*. London : Longman.

Freebody, P. y Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly* 18 (3): 277-294.

Fulcher, F. y Davidson, F. (2009). Test architecture, test retrofit. *Language Testing* 26 (1) 123-144.

Fulcher, G. (1999a). Assessment in English for Academic Purposes: Putting Content Validity in its Place. *Applied Linguistics* 20 (2): 221-236.

Fulcher, G. (1999b). Ethics in Language Testing. TAE SIG Newsletter 1 (1): 1-3. (Documento disponible en Internet en: http://taesig.8m.com/news1.html).

García Gómez, P., Noah, A., Schedl, M., Wright, C. y Yolkut, A. (2007). Proficiency descriptors based on a scaled-anchoring study of the new TOEFL iBT reading tests. *Language Testing* 24 (3): 417-444.

Gillham, B. (2000). *Developing a questionnaire*. London: Continuum.

Grabe, W. (1991). Current developments in second-language reading research. *TESOL Quarterly* 25 (3): 375-406.

Grabe, W. (2000). Developments in reading research and their implications for computer-adaptive reading assessment. En M. Chalhoub-Deville (eds.), *Issues in computer-adaptive tests of reading*. Cambridge: Cambridge University Press.

Graña López, B. (1997) Frecuencia y procesamiento léxico. Revista Española de lingüística Aplicada 12: 27-41

Green, A. B y Weir, C. J. (2004). Can placement test inform instructional decisions?. *Language Testing* 21 (4): 467-494.

Grotjahn, R. (1987). How to construct and evaluate a C-test: a discussion of some problems and some statistical analysis. En Rüdiger Grothan, Christine Klein-Braley & Douglas K. Stevenson (eds.), *Taking their measure: The validity and validation of language tests*, 219-253. Bochum: Brockmeyer.

Grotjahn, R. (2006). *The C-Test: Theory, Empirical Research, Applications*. Frankfurt: Peter Lang.

Grotjahn, R. y Stemmer, B. (2002). C-tests and language processing. En James A. Coleman, Rüdiger Grotjan & Ulrich Raatz (eds.), *University language testing and the C-test* (115-130). Bochum: AKS-Verlag.

Hadley, G. y Naaykens, J. E. (2006). An Investigation of the Selective Deletion Cloze test as a Valid Measure of Grammar-Based Proficiency in Second Language Learning. (Documento disponible en Internet en: http://www.nuis.ac.jp/~hadley/publication/nucloze/NUCLOZE.htm).

Hale, G. A. (1988). Student major field and text content : interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing* 5 (1): 46-61.

Hamp-Lyons, L. (1997). Washback, impact and validity: ethical concerns. Language Testing 14 (3): 295-303.

Hanson, F. A. (1993). *Testing: Social Consequences of the Examined Life*. Berkeley, CA: University of California Press.

Harrison, A. (1991). Language Assessment as Theatre: Ten Years of communicative Testing. En J. C. Alderson y B. North, *Language Testing in the 1990s: The Communicative Legacy* , 95-105. London: Mcmillan.

Hawthorne, L. (1997). The political dimension of English language testing in Australia. *Language Testing* 14 (3): 248-260.

Heaton, J. B. (1975). *Writing English Language Tests*. London: Longman.

Heaton, J. B. (1988). *Writing English Language Tests*. London: Longman.

Heaton, J. B. (1990). *Classroom Testing*. New York: Longman.

Heilenman, L. K. (1983). The Use of a Cloze Procedure in Foreign Language Placement. *The Modern Language Journal* 67: 121- 126.

Henning, G. (1987). *A guide to language testing*. Cambridge, Mass: Newbury House.

Herrera Soler, H. (1999). Is the English test in the Spanish University Entrance Examination as discriminating as it should be? *Estudios Ingleses de la Universidad Complutense* 7: 87-103.

Herrera Soler, H. (2001). "Clozes" prototípicos en la evaluación de la comprensión lectora. En Ana I. Moreno y Vera Colwell (eds.) *Perspectivas Recientes sobre el discurso.* Universidad de León. CD:1- 9.

Herrera Soler, H. et al. (2003). An analysis of a multiple choice test for a personnel selection process. En Gloria Luque et al (eds). Las Lenguas en un mundo global: Languages in a global world. CD, 135-149.

Herrera Soler, H. (2004). Interpretación de los resultados de una prueba de elección múltiple. En Marta Carretero et al. (eds.). Estudios de Lingüística Aplicada a la Comunicación. C.E.R.S.A: Madrid. 89 – 108.

Herrera Soler, H. (2005). El Test de Elección Múltiple: Herramienta Básica en la Selectividad. En H. Herrera Soler y J. García Laborda. Estudios y Criterios para una Selectividad de Calidad en el Examen de Inglés. Universidad Politécnica de Valencia: Valencia. 65 – 98.

Herrera Soler, H y Martínez Arias, R. (2000). Influence of the scoring procedure on reliability and validity of an achievement test. *Proceeding of the International Conference on Measurement and Multivariate Analysis*l. ICMMA. Banff, Alberta, Canada, 161- 166.

Herrera Soler, H. y Martínez Arias, R. (2002). A new insight into examinee behaviour in a multiple-choice test: a quantitative approach. Estudios Ingleses de la Universidad Complutense. V. 10: 113-137.

Herrera Soler, H, Amengual, M y Esteban, M. (2001) Lectura de la prueba de inglés de la selectividad desde una perspectiva pitagórica. En Carmen Valero et al (eds.) *Lingüística Aplicada a Finales del Siglo XX*. Universidad de Alcalá de Henares. Volumen: 1, 177-184.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoey, M. (2001). *Textual Interaction*. Oxford: Oxford University Press.

Hoey, M. (2005). *Lexical priming: a new theory of words and language*. London:Routledge

Holliday, A y Cook, T. (1982). An ecological approach to ESP. Lancaster Practical Papers in English Language Education. *Issues in ESP* 5: 123-143.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Huhta, A. et al. (2006). Discursive construction of a high-stakes test: the many facets of a test-taker. *Language Testing* 23 (3): 326-350.

Huitt, B. et al. (2001). *Assessment, Measurement, Evaluation and Research*. (Documento de Internet disponible en: www.adprima.com/measurement.htm).

Hulstijn, J. H. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *The Modern Language Journal* 91: 663-667.

Hutchinson, T. y Waters, A. (1987). *English for Specific Purposes*. Cambridge: Cambridge University press.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4-21.

Ikeguchi, C. B. (1998). Do different C-tests discriminate proficiency levels of EL2 learners?. JAALT Testing and Evaluation SIG Newsletter 2 (1): 2-10. (Documento de Internet disponible en: http://www.jalt.org/test/ike_2.htm).

ILTA (2000). Code of Ethics for ILTA. (Documento de Internet disponible en: http://www.Dundee.ac.uk/languagestudies/1test/ilta/ilta_test2.html).

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing* 25 (3): 385-402.

Jafarpur, A. (1995). Is C-testing superior to cloze?. *Language Testing* 12 (2): 194-214.

Jafarpur, A. (1999a). Can the C-test be improved with classical item analysis?. *System* 27: 79-89.

Jafarpur, A. (1999b). What's Magical About the Rule of Two for constructing C-Tests*?. RECL Journal* 30 (2): 86-100.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing* 26 (1): 31-73.

Jennings, M., Fox, J., Graves, B. y Shohamy, E. (1999). The test-takers´ choice: an investigation of the effect of topic on language test performance. *Language Testing* 16 (4): 422-456.

Johnson, J. S. y Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive psychology* 21: 60-69.

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly* 19 (2): 219-239.

Jones, C. (1991). An integrated model for ESP syllabus design. *English for Specific Purposes* 10 (3):155-172.

Jones, R. L. y Spolsky, B. (eds.), (1975). *Testing Language Proficiency*. Arlinton, Va.: Center for applied Linguistics.

Jonz, J. (1976). Improving on the basic egg: The m-c cloze*. Language Learning* 26 (2): 255-265.

Jonz, J. (1990). Another Turn In The Conversation. What Does Cloze Measure?. *TESOL Quarterly* 24 (1): 61-83.

Jonz, J. (1991). Cloze item types and second language comprehension. *Language Testing* 8 (1): 1-22.

Katona, L. y Dörnyei, Z.(1993). The C-Test. *Forum* 31 (2): 35-38.

Kitano, K. (2001). Anxiety in the college Japanese language classroom. *Modern Language Journal* 85 (4): 549-566.

Kizlik, B. (2007). *Measurement, Assessment, and Evaluation in Education*. (Documento de Internet disponible en: www.adprima.com/measurement.htm).

Klein-Braley, C. (1981). *Empirical Investigations of Cloze Tests*. Doctoral dissertation. University of Duisburg.

Klein-Braley, C. (1984). Advance Prediction of Difficulty with C-test. En T. Culhane et al. (eds.), *Practice and problems in language testing*. Occasional papers 29: 97-112. Colchester: University of Essex.

Klein-Braley, C. (1985). A cloze-up on the C-Test: a study in the construct validation of authentic tests. *Language Testing* 2: 76-104.

Klein-Braley, C. (1994). Language Testing with the C-Test. A Linguistic and Statistical Investigation into the Strategies Used by C-Test Takers, and the prediction of C-test Difficulty. Habilitationsschrift, University of Duisburg.

Klein-Braley, C. (1997). C-test in the context of reduced redundancy testing: an appraisal. *Language Testing* 14 (1): 47-84.

Klein-Braley, C. (2002). Psycholinguistics of C-test Taking. En James A. Coleman, Rüdiger Grotjan & Ulrich Raatz (eds.), *University language testing and the C-test*, 131-142. Bochum: AKS-Verlag.

Klein-Braley, C. y Raatz, U. (1984). A survey of research on the C-test. *Language testing* 1: 134-146.

Klein-Braley, C. y Raatz, U. (1998). Introduction to language testing and C-tests. Universidad de Duisburg. (Documento de Internet disponible en: www.uni-Duisburg.De/fb3/angling/forschung/howtodo.htm).

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing* 19 (2): 193-220.

Koda, K. (1997). Orthographic knowledge in L2 lexical processing: A cross-linguistic perspective. En J. Coady y T. Huckin (eds.), *Second Language Vocabulary Acquisition* (35-52). Cambridge: Cambridge University Press.

Koda, K. (2005). Insights into Second Language Reading. Cambridge: Cambridge University Press.

Kokkota, V. (1988). Letter-deletion procedure: a flexible way of reducing text redundancy. *Language Testing* 5 (1): 115-119.

Kontra, E. H. y Kormos, J. (2006). Strategy use and the construct of C-tests. En R. Grotjahn, *The C-Test: Theory, Empirical Research, Applications*, 121-138. Frankfurt: Peter Lang.

Krashen, S. D., Long, M. H. y Scarcella, R. C. (1979) age rate and eventual attainment in second language acquisition. *TESOL Quarterly* 13: 573-582.

Kroll, F. et al. (2002). The development of lexical fluency in a second language. *Second Language Research* 18 (2): 137-171.

Krumm, H. J. (2007). Profiles Instead of Levels: The CEFR and Its (Ab) Uses in the Context of Migration. *The Modern Language Journal* 91: 667-669.

Laufer, B. (1992). How much lexis is necessary for reading comprehension. En Arnaud and Béjoint, 126-132.

Laufer, B. (1997a). What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, 140-155. Cambridge: Cambridge University Press.

Laufer, B. (1997b). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. En J. Coady y T. Huckin (eds.), *Second language Vocabulary Acquisition*, 20-34. Cambridge: Cambridge University Press.

Laufer, B. (1998). The development of passive and active vocabulary; same or different?. *Applied Linguistics* 19: 255-271.

Laufer, B. y Goldstein, Z. (2004). Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning* 54 (3): 399-436.

Laufer, B. y Nation, P. (1995). Vocabulary Size and Use: Lexical richness in L2 Written Production. *Applied Linguistics* 16 (3): 307-322.

Laufer, B. y Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing* 16 (1): 33-51.

Laufer, B. y Paribakht, T. S. (1998). The Relationship Between Passive and Active Vocabularies: Effects of Language Learning Context. *Language Learning* 48 (3): 365- 391.

Laufer, B., Elder, C., Hill, K. y Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge?. *Language testing* 21 (2): 202-226.

Lawson, M. J. y Hogben, D. (1996). The Vocabulary-Learning Strategies of Foreign-Language Students. *Language learning* 46 (1): 101-135.

Lee, S. (!996). The Concurrent Validity of Cloze Tests with Essay Tests Among Korean Students. *Texas Papers in Foreign Language* 2 (2): 57-69.

Lee, S. (2007). Effects of Textual Enhancement and Topic Familiarity on Korean EFL Students´ Reading Comprehension and Learning of passive Form. *Language Learning* 57 (1): 87-118.

Lee, S. H. (2003). ESL learner's vocabulary use in writing and the effects of explicit vocabulary instruction. *System* 31:537-561.

Lee, S. H. (2008). Beyond reading and proficiency assessment: The rational Cloze procedure as stimulus for integrated reading, writing, and vocabulary instruction and teacher-student interaction in ESL. *System* 36: 642-660.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an ELF reading comprehension test. *Language Testing* 21 (1): 74-100.

Lee, Y. P. (1985). Investigating the Validity of the Cloze Score. En *New Directions in Language Testing* 237-147. Lee et al., (eds.) Oxford: Pergamon Press.

Lee, Y. P. et al. (eds.), (1985). *New Directions in Language Testing*. Papers presented at the International Symposium on Language Testing. Hong Kong. Oxford: Pergamon Press.

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rash Analysis. *Language testing* 26 (2): 245-274.

Lenneberg, E. H. (1967). *Biological Foundations of Language*. New York: Wiley

Lewkowick. J y Moon, J. (1985). Evaluation: A Way of Involving the Learner. En J. C. Alderson (eds.), Lancaster practical Papers in English Language Education. Vol.6. *Evaluation* (45-80). Oxford: Pergamon Press.

Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing* 17 (1): 43-64.

Little, D. (2002). The European Language Portfolio: structure, origins, implementation and challenges. *Language teaching* 35: 182-189.

Little, D. (2005). The Common European Framework and the European Language Portfolio: involving learners and their judgements in the assessment process.. *Language Testing* 22 (3): 321-336.

Little, D. (2006). The Common European Framework of Reference for Languages: content, purpose, origin, reception and impact. *Language teaching* 39: 167-190.

Little, D. (2007). The common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. *The Modern Language Journal* 91: 645-655.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing* 24 (4): 489-515.

Low, G. (1985). Validity and the problems of direct language proficiency tests. En J. C. Alderson (eds.). Lancaster practical Papers in English Language Education. Vol.6. *Evaluation*, 151-168. Oxford: Pergamon Press.

Lumley, T. y O´Sullivan, B. (2005). The effect of test–taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing* 22 (4): 415-437.

Lynch, B. K. (1997). In search of the ethical test. *Language Testing* 14 (3): 315-327.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing* 18 (4): 351-372.

Lyons, J. (1981). *Language and Linguistic*. Cambridge: Cambridge University Press.

Machida, S. (2001). Anxiety in Japanese-language class oral examinations *Sekai no Nihongo Kyoiku* 11: 115-138.

MacIntyre, P. D., Baker, S. C., Clément, R. y Donovan, L. A. (2002). Sex and age effects on willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high school French immersion students. *Language Learning* 52 (3): 537-564.

Mackay, R. (1978). Identifying the nature of learners´ needs. En R. Mackay y A. Mountford (eds.), *English for Specific Purposes* 21-42. London: Longman.

Mackey, A. y Gass, S. M. (2005). *Second Language Research: Methodology and Design*. Mahwah, New Jersey y London: Lawrence Erlbaum Associates.

Madaus, G. (1990). *Testing as a Social Technology. The Inaugural Annual Boise Lecture on Education and Public Policy*, Boston, MA: Boston College.

Matsuda, S. y Gobel, P. (2004). Anxiety and predictors of performance in the foreign language classroom. *System* 32:21-36.

Mauranen, A. (1989). Can Gaps Measure Comprehension?. Modifications of Cloze as Tests of Reading. In: *Special Language. From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters Ltd, 337-346.

McBeath, N. (1989). C-Tests in English: Pushed beyond the Original Concept?. *RELC Journal* 20 (2): 36-41.

McCarthy, M. y Carter, R. (1997). Written and spoken vocabulary. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy* (20-39), Cambridge: Cambridge University Press.

McCarthy, P. M. y Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing* 24 (4): 459-488.

McNamara, T. y Roever, C. (2006). *Language testing: The social dimension*. Malden, MA y Oxford: Blackwell.

McNamara, T. (1997). *Measuring Second Language performance*. London y New York: Longman.

McNamara, T. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics* 18: 304-319.

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

McNamara, T. (2001a). Rethinking alternative assessment. *Language Testing* 18 (4): 329-332.

McNamara, T. (2001b). Language assessment as social practice: challenges for research. *Language Testing* 18 (4) 333-349.

McNamara, T. (2003). Looking back, looking forward: rethinking Bachman. *Language Testing* 20 (4): 466-473.

McNamara, T. (2007). Language Assessment in Foreign Language Education: The Struggle over Constructs. *The Modern Language Journal* 91: 280-283.

Meara, P. (1990). A note on passive vocabulary. *Second Language Research* 6: 150-154.

Meara, P. (1992). Network Structures y Vocabulary acquisition in a Foreign Language. En Arnaud, J. L. y Béjoint, H. (eds.), (1992). *Vocabulary and Applied Linguistics*, 62-71 London: Mcmillan.

Meara, P. y Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing* 4: 142-151.

Meara, P. (1997). Towards a new approach to modelling vocabulary. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy* (109-121), Cambridge: Cambridge University Press.

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research* 18 (4): 393-407.

Meara, P. (2005). Lexical Frequency Profiles: A Monte Carlo Analysis. *Applied Linguistics* 26 (1): 32- 47.

Meara, P. y Fitzpatrick, T. (2000). Lex30: an improved method of assessing productive vocabulary in an L2. *System* 28: 19-30.

Mejías, H., Applebaum, R. L., Applebaum, S. J. y Trotter, S. J. (1991). Oral communication apprehension and Hispanics: an exploration of oral communication apprehension among Mexican American Students in Texas. En Horwitz, E. K. Y young, D. J. (eds.), *Language Anxiety: From Theory and Research to Classroom Implications,* 87-97. Englewood Cliffs, NJ: Prentice Hall.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy* (84-102), Cambridge: Cambridge University Press.

Messick, S. (1981). Evidence and Ethics in the Evaluation of tests. *Educational Researcher* 10: 9-20.

Messick, S. (1989). Validity. In Linn, R. L., editor*, Educational measurement* (13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23: 13-23.

Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13 (4): 241-257.

Mochida, A. y Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing* 23 (1): 73-98.

Moon, R. (1997). Vocabulary connections: multi-word items in English. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy* (40-63), Cambridge: Cambridge University Press.

Murphy, D. F. (1985). Evaluation in language teaching: assessment, accountability and awareness. En Lancaster practical Papers in English Language Education. Vol.6. *Evaluation*, 1-18. Oxford: Pergamon Press.

Nagy, W. (1997). On the role of context in first and second language vocabulary learning. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy* (64-83), Cambridge: Cambridge University Press.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. New York: Heinle and Heinle

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P. y Kiongho, H. (1995). Where would general service vocabulary stop and special purposes vocabulary begin?. System 23 (1): 35-41.

Nation, P. (1983). Testing and Teaching Vocabulary. *Guidelines* 5: 12-25.

Nation, P. y Coady, J. (1988). Vocabulary and Reading. En R. Carter Y. M. McCarthy (eds.), *Vocabulary and language teaching* (97-110). London: Longman.

Nation, P. y Waring, R. (1997). Vocabulary size, text coverage and word lists. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy* (6-20), Cambridge: Cambridge University Press.

Nattinger, J. (1988). Some current trends in vocabulary teaching. En R. Carter y M. McCarthy (eds.), *Vocabulary and language teaching* (62-82). London: Longman.

Nattinger, J. R. y DeCarrico, J. S. (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics* 24 (2): 223 -242.

Neuner, G. (1992). The Role of Experience in a Content-and-Comprehension-Oriented Approach to Learning a Foreign Language. En Arnaud, J. L. y Béjoint, H. (eds.), (1992). *Vocabulary and Applied Linguistics*, 156-167. London: Mcmillan

Norris, J.M. (2006). Development and evaluation of a curriculum-based German C-test for placement purposes. En R. Grotjahn, *The C-Test: Theory, Empirical Research, Applications* (45-84). Frankfurt: Peter Lang.

North, B. (2007). The CEFR Illustrative Descriptor Scales. *The Modern Language Journal* 91: 256-259.

Nunan, D. (1988). *The Learner-Centred Curriculum*. Cambridge: Cambridge University Press.

Nunan, D. (1992). *Research Methods in Language Learning*. Cambridge: Cambridge University Press.

Odlin, T. y Natalico, D. (1982). Some characteristics of word classification in a second language. *The Modern Language Journal* 66: 34-38.

O´Malley, M. J. y Chamot, A. U. (1990). *Learning strategies in second language acquisition.* Cambridge: Cambridge University Press.

O´Neil, K. A., McPeek, W. M. y wild, C. L. (1993). Differential Item functioning on the Graduate Management Admission Test*. ETS Research Report* 35-93. New Yersey.

Oller, J. W. (1973). Cloze tests and second language proficiency and what they measure. *Language learning* 23 (1).

Oller, J. W. Jr. (1979). *Language Tests at School*. London & New York: Longman.

Oller, J. W. Jr. (1983). *Issues in language testing research*. Rowley, Massachusetts: Newbury House.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter.

Oxford, R. L. (1990). *Language Learner Strategies*: What every teacher should know. New York: Newbury House.

Pae, T. (2004). Gender effect on reading comprehension with Korean EFL learners. *System* 32: 265-281.

Penfield, W. y Roberts, L. (1959). Speech and Brain Mechanisms. New York: Atheneum.

Perkins, K. (1992). The effect of passage topical structure types on ESL reading comprehension difficulty. *Language testing* 9 (2): 163-173.

Perkins, K. y Brutten, S. R. (1988). An item discriminability study of textually explicit, textually implicit, and scriptally implicit questions. *RELC Journal*. 19 (2): 1-11.

Phakiti, A. (2003). A Closer Look at Gender and Strategy Use in L2 Reading. *Language Learning* 53 (4): 649-702.

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing* 25 (2): 237-272.

Poehner, M. E. (2007). Beyond the Test: L2 Dynamic Assessment and the Transcendence of mediated Learning. *The Modern Language Journal* 91: 323-340.

Pollitt, A. (1991a). Giving Students a Sporting Chance: assessment by Counting and by Judging. En J. C. Alderson y B. North, *Language Testing in the 1990s: The Communicative Legacy*, 46-59. London: Mcmillan.

Pollitt, A. (1991b). Response to Charles Alderson´s Paper: "Bands and Scores". En J. C. Alderson y B. North, *Language Testing in the 1990s: The Communicative Legacy* 87-94. London: Mcmillan.

Porter, D. (1991). Affective Factors in Language Testing. En J. C. Alderson y B. North, *Language Testing in the 1990s: The Communicative Legacy* 32-40. London: Mcmillan.

Potts, P. J. (1985). The Role of Evaluation In a communicative Curriculum, and some Consequences for Materials Design. En J. C. Alderson (eds.), Lancaster practical Papers in English Language Education. Vol.6. *Evaluation*, 19-44. Oxford: Pergamon Press.

Pritchard, R. H. (1990). The effects of cultural squemata on reading processing strategies. Reading Research Quarterly 25: 273-295.

Prodomou, L (1995) the Backwash Effect: from Testing to Teaching. *Language Testing* 49 (1): 13-25.

Pulido, D. (2003). Modelling the Role of Second Language Proficiency and topic Familiarity in second Language Incidental Vocabulary Acquisition Through Reading. *Language learning* 53 (2): 233-284.

Pulido, D. (2007). The Effects of Topic Familiarity and Passage Sight Vocabulary on L2 Lexical Inferencing and Retention through Reading. *Applied Linguistics* 38 (1):66-86.

Pulido, D. (2007). The Relationship Between Text Comprehension and Second Language Incidental Vocabulary Acquisition: A Matter of topic Familiarity?. *Language Learning* 57 (1): 155-199.

Qian, D. D. (1999). Assessing the roles of depth and breath of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review* 56: 282-308.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective. *Language Learning* 52: 513-536.

Qian, D. D. (2007). Assessing University Students: Searching for a English Language Exit Test. *Regional Language Centre Journal* 38 (1): 18-37.

Qian, D. D. y Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing* 21 (1): 28-52.

Raatz, U. (1984). The factorial validity of C-Tests. En T. Culhane et al. (eds.), (1984). *Practice and problems in language testing*. *Occasional papers* 29, 124-139. Colchester: University of Essex.

Raatz, U. (1985). Better theory for better tests?. *Language Testing* 2 (1): 60-75.

Raatz, U. (2002). Introduction to language testing and C-Tests. Universidad de Duisburg. (Documento de Internet disponible en: www.uni.duisburg.de/fb3/angling/forschung/howtodo.htm).

Raatz, U. y Klein-braley, C. (1982). The C-test: a modification of the cloze procedure. En Culhane et al., (eds.). (1982).

Rashid, S. M. *Validating the C-test amongst Malay ESL Learners*. (Documento de Internet disponible en: www.melta.org.my/modules/sections/12.doc).

Rea Dickins, P. (2001). Mirror, mirror on the wall: identifying processes of classroom assessment. *Language Testing* 18 (4): 429-462.

Rea Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing* 21 (3): 249-258.

Rea Dickins, P. M. (1991). Response to Andrew Harrison's Paper: "Language Assessment as Theatre". En J. C. Alderson y B. North, *Language Testing in the 1990s: The Communicative Legacy* (106-111). London: Mcmillan.

Rea, P. M. (1984). Language tests as indicators of academic achievement. En Culhane, T. et al. (eds.), *Practice and problems in language testing*. Occasional papers 29, 140-158. Colchester: University of Essex.

Read, J. (1997). Vocabulary and testing. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy* 303-320, Cambridge: Cambridge University Press.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Read, J. y Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing* 18 (1): 1-32.

Richterich, R. (1973/1980). Definition of Language needs and types of adults. En J. R. Trim, D. Van Ek, Y D. Wilkins (eds.) *Systems Development in Adult Language Learning* 29-88. Oxford/Strasbourg: Pergamon/Council of Europe.

Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly* 10: 77-89.

Richterich, R. (1983). *Case Studies in Identifying Language Needs*. Oxford: Pergamon/Council of Europe.

Riley, G. L. Y Lee, J. E. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing* 13 (2): 173-189.

Rimmer, W. (2006). Measuring grammatical complexity: the Gordian knot. *Language testing* 23 (4): 497-519.

Ross, S. (1998). Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. *Language testing* 15 (1): 1-20.

Rupp, A. A., Ferne, T. y Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing* 23 (4): 441-474.

Ryan, A. (1997). Learning the orthographic form of L2 vocabulary – a receptive and productive process. En N. Schmitt y M. McCarthy (eds.) *Vocabulary: Description, Acquisition and Pedagogy* (181-198), Cambridge: Cambridge University Press.

Salager-Meyer, F. (1991). Reading expository prose at the post-secondary level: the influence of textual variables on L2 reading comprehension (a genre-based approach). *Reading in a Foreign Language* 8 (1): 645-662.

Sarig, G. (1989). Testing meaning construction: can we do it fairly?. *Language Testing* 6 (1):77-94.

Sasaki, M. (2000). Effects of cultural schemata on students´ test-taking processes for cloze tests: a multiple data source approach. *Language Testing* 17 (1): 85-104.

Scarcella, R. y Zimmerman, C. (1998). Academic words and gender: ESL student performance on a test of academic lexicon. *Studies in Second Language Acquisition* 20: 27-49.

Schmitt, N. (1997). Vocabulary learning strategies. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, 199-227. Cambridge: Cambridge University Press.

Schmitt, N. (1998). Quantifying word association responses: what is native-like?. *System* 26 (3): 389-401.

Schmitt, N. (1999). The relationship between TEOFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing* 16 (2) 189-216.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University press.

Schmitt, N. y McCarthy, M. (eds.), (1997). *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D. y Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. Language Testing 18 (1): 55-88.

Schoonen, R. y Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing* 25 (2): 211-236.

Shiotsu, T y Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing* 24 (1): 99-128.

Shohamy, E. (1998). How can language testing and SLA benefit from each other?. The case of discourse. En L. F. Bachman y A. D. Cohen. *Interfaces Between Second Language Acquisition and Language Testing Research*, 156-176. Cambridge: Cambridge University Press.

Shohamy, E. (2000). The relationship between language testing and second language acquisition, revisited. *System* 28: 542-553.

Shohamy, E. (2001a). *The power of tests. A critical perspective on the uses of language tests*. Harlow, Essex: Pearson Education.

Shohamy, E. (2001b). Democratic assessment as an alternative. *Language Testing* 18 (4): 373-391.

Shohany, E. (1984a). Does the testing method makes a difference?. The case of reading comprehension. *Language testing* 1(2): 147-170.

Shohany, E. (1984b). Input and output in language testing. En T. Culhane et al. (eds.), *Practice and problems in language testing*. Occasional papers 29, 159-176. Colchester: University of Essex.

Sigott, G. (2004). *Towards Identifying the C-test Construct.* Frankfurt: Peter Lang.

Sigott, G. (2006). How fluid is the C-Test construct?. En R. Grotjahn, *The C-Test: Theory, Empirical Research, Applications*, 139-146. Frankfurt: Peter Lang.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Singleton, D. (1999). *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.

Singleton, D. y Little, D. (1991). The second language lexicon: some evidence from university-level learners of French and German. *Second Language Research* 7: 61-68.

Singleton, D. y Singleton E. (2002). The C-test and L2 lexical acquisition/ processing research. En James A. Coleman, Rüdiger Grotjan & Ulrich Raatz (eds.), *University language testing and the C-test*, 143-168. Bochum: AKS-Verlag.

Skehan, P. (1991). Progress in Language Testing: the 1990s. En J. C. Alderson y B. North, *Language Testing in the 1990s: The Communicative Legacy*, 3-21. London: Mcmillan.

Snellings, P. et al. (2004). Validating a test of second language written lexical retrieval: a new measure of fluency in written language production. *Language testing* 21 (2): 174-201.

Sökmen, A. J. (1997). Current trends in teaching second language vocabulary. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, 237-257. Cambridge: Cambridge University Press.

Spence-Brown, R. (2001). The eye of the beholder: authenticity in an embedded assessment task. *Language testing* 18 (4): 463-481.

Spolsky, B. (1973). What does it mean to know a language; or how do you get someone to perform his competence?. En J. Oller, y J. Richards (eds). *Focus on the learner*, 164-176. Rowley Massachusetts: Newbury House Publishers.

Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing*. Oxford: Oxford University Press.

Spolsky, B. (1997). The ethics of gate keeping tests: what have we learned in a hundred years?. Language Testing 14 (3): 242-247.

Spolsky, B. (1981). Some ethical questions about language testing. En C. Klein-Braley, y D. K. Stevenson. (eds.). (1981). Practice and problems in language testing 1, 5-21. Frankfurt: Lang.

Spolsky, B., Bengt, S., Sato, M., Walker, E. y Aterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. *Language Learning Special Issue* 3: 79-101.

Stevenson, D. K. (1985). Pop Validity and Performance Testing. En Y. P. Lee et al. (eds.), *New Directions in Language Testing*. Papers presented at the International Symposium on Language Testing. Hong Kong. Oxford: Pergamon Press.

Storey, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse cloze test. *Language Testing* 14 (2): 214-231.

Süssmilch, E. (1984). Language testing with immigrant children. En T. Culhane et al. (eds.), *Practice and problems in language testing*. Occasional papers 29: 167-176. Colchester: University of Essex.

Swan, M. (1997). The influence of the mother tongue on second language vocabulary acquisition and use. En N. Schmitt y M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, 156-180. Cambridge: Cambridge University Press.

Tadamitsu, K. (2001). An examination of Nation's (1990) Vocabulary Levels Test. (Documento de Internet disponible en: http://www.1.harenet.ne.jp/~waring/vocab/colloquium/tad2001.htm)

Tarone, E y Yule, G. (1989). *Focus on the Language Learner*. Oxford: Oxford University Press.

Tarone, E. (1998). Research on interlanguage variation: Implications for language testing. En L. F. Bachman, y A. D. Cohen, *Interfaces Between Second Language Acquisition and Language Testing Research*, 71-89. Cambridge: Cambridge University Press.

Taylor, W. L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly* 30: 415-453.

Thomas, M. (1994). Assessment of L2 Proficiency in Second Language Acquisition Research. *Language Learning* 44 (2): 307-336.

Unquhart, A. H. y Weir, C. J. (1998). *Reading in a Second Language: Process, Product and Practice*. Harlow: Longman.

Valdés, G. y Figueroa, R. (1996). *Bilingualism and Testing: A Special Case of Bias*. Norwood, NJ: Ablex Publishing Corp.

Van Dijk, T. A. (1977). *Text and Context: Explorations in the Semantics and Pragmatics of discourse*. London: Longman.

Vann, J. R. y Abraham, G. R. (1990). Strategies of unsuccessful language learners. *TESOL Quarterly* 24: 177-198.

Wall, D. (1996). Introducing new tests into traditional systems: insights from general education and from innovation theory. *Language Testing* 13 (3): 334-335.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: can this be predicted or controlled?. *System* 28: 499-509.

Wall, D. y Alderson, C. (1996). Examining Washback: The Sri Lankan Impact Study. In A. Cumming y R. Berwick (eds.) *Validation in Language Testing*. Clevedon: Multilingual Matters Ltd. (194-221).

Waring, R. (1998). Receptive and Productive Foreign Language Vocabulary Size II. (Documento en Internet disponible en: www.fltr.ucl.ab.bc/fltr/germ/etan/bibs/vocab/RPII.html).

Weir , C. (1988). *Communicative language testing*. Exeter: University of Exeter

Weir, C. (1990). *Communicative Language testing*. Hemel Hempstead: Prentice-Hall.

Weir, C. (1993). *Understanding & developing Language Tests*. Hemel Hempstead: Prentice Hall International.

Weir, C. J. (2005a). *Language Testing and Validation: An Evidence-Based Approach*. New York: Palgrave Macmillan.

Weir, C. J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing* 22 (3): 281-300.

Weir, C. J. y Roberts, J. R. (1994). *Evaluation in ELT*. Oxford: Blackwell.

Wen, Q. y Johnson, R. K. (1997). L2 learner variables and English achievement: a study of tertiary-level English majors in China. *Applied Linguistics* 18: 27-48.

Wode, H., Rohde, A., Gassen, F., Weiss, B., Jekat, M. y Jung, P. (1992). L1, L2, L3: Continuity vs. Discontinuity in Lexical Acquisition. En Arnaud, J. L. y Béjoint, H. (eds.), (1992). *Vocabulary and Applied Linguistics*, 52-62. London: Mcmillan

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *Modern Language Journal* 77 (4): 473-489.

Wolter, B. (2002). Assessing proficiency through word associations: is there still hope?. *System* 30: 315-329.

Wood, R. (1991). *Assessment and Testing: A Survey or Research*. Cambridge: Cambridge University Press.

Wood, R. (1993). *Assessment and Testing*. Cambridge: Cambridge University Press.

Woods, A., Fletcher, P. y Hughes, A. (1986). *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Yamada, K. (2005). Lexical Patterns in the Eyes of Intermediate EFL Readers. *Regional Language Centre Journal* 36 (2): 177-188.

Yamada, K. (2009). Lexical patterns in L2 textual gist identification assessment. *Language Testing* 26 (1): 101-122.

Yamashita, J. (2003). Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers. *Language Testing* 20 (3): 267-293.

Young, D. J. y Oxford, R. (1997). A gender-related analysis of strategies used to process writing input in the native language and a foreign language. *Applied Language Learning* 8: 43-73.

Zareva, A. (2005). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System* 33: 547-562.

Zeidner, M. (1998). *Test anxiety: the state of the art*. New York y London: Plenum Press.

Zimmerman, C. B. (1997). Historical trends in second language vocabulary instruction. En J. Coady y T. Huckin. (eds.), *Second language Vocabulary Acquisition*, 5-19. Cambridge: Cambridge University Press.

BASES DE DATOS CONSULTADAS

BNC (British National Corpus).: http://info.ox.ac.uk/bnc/

Brown Corpus: http:// homepages.infoseek.com/~corpuslinguistics/

PROGRAMAS INFORMÁTICOS

SPSS (Statistical Package for the Social Sciences) 15.0 for windows

APÉNDICES

APÉNDICE 1: C-TESTS

C-Test A

NAME……...

Directions: The following test has been developed by removing the second half of every second word in four different texts beginning with the second sentence. The missing part contains the same number of letters as the first part or one more letter than the first part. No contracted forms have been used, no proper names or numbers have been deleted. You are supposed to reconstruct the texts.

Example:

A British university is now doing research into the difference between men and women drivers. It se_ _ _ that wo_ _ _ often dr_ _ _ more care_ _ _ _ _ than m_ _.

A British university is now doing research into the difference between men and women drivers. It **seems** that **women** often **drive** more **carefully** than **men.**

PHYSICAL EXERCISE

We all need exercise, especially young people in their teens and adults from twenty to eighty. Regular exer_ _ _ _ (1) temporally ti_ _ _ (2) the bo_ _ (3) but la_ _ _ (4) on actu_ _ _ _ (5) gives y_ _ (6) more ene_ _ _ (7). This i_ (8) why peo_ _ _ (9) who suf_ _ _ (10) from gen_ _ _ _ (11) tiredness c_ _ (12) benefit fr_ _ (13) taking mo_ _ (14) exercise rat_ _ _ (15) than mo_ _ (16) rest. Exer_ _ _ _ (17) makes y_ _ (18) feel a_ _ (19) look bet_ _ _ (20) and c_ _ (21) also he_ _ (22) you t_ (23) lose wei_ _ _ (24) because i_ (25) burns up fat or food to produce energy. However, if you are over 40, or if you have recently had a serious illness, you should visit your doctor before starting a general exercise routine.

RELAX AND LIVE

It is commonly believed that only rich middle-aged businessmen suffer from stress, but in fact anyone may become ill as a result of stress if they experience a lot of worry over a long period and their health is not particularly good. Stress c____ (1) be a fri____ (2) or a____ (3) enemy: i____ (4) can wa____ (5) you th____ (6) you a____ (7) under t____ (8) much pres____ (9) and sho____ (10) change yo____ (11) way o____ (12) life. I____ (13) can ki____ (14) you i____ (15) you d____ (16) not not____ (17) the war____ (18) signals. Doc____(19) agree th____ (20) it i____ (21) probably t____ (22) biggest ca____ (23) of ill____ (24) in t____ (25) western world. Stress can cause car accidents, heart attacks, alcoholism and may even drive people to suicide. As with all illnesses prevention is better than cure so if you are not able to relax then it is time to stop and ask yourself whether your present life really suits you.

ALTERNATIVE SOURCES OF ENERGY

The world is running out of oil and governments are searching for a suitable alternative, but so far in vain. Many scientists a_ _ (1) optimistic th_ _ (2) new wa_ _ (3) of gener_ _ _ _ _ (4) large amo_ _ _ _ (5) of ene_ _ _ (6) will b_ (7) successfully deve_ _ _ _ _ (8), but a_ (9) the sa_ _ (10) time th_ _ (11) fear t_ _ (12) consequences. I_ (13) the wo_ _ _ (14) population go_ _ (15) on incre_ _ _ _ _ (16) at i_ _ (17) present ra_ _ (18), and ea_ _ (19) individual conti_ _ _ _ (20) to u_ _ (21) more ene_ _ _ (22) every ye_ _ (23), we wi_ _ (24) damage t_ _ (25) earth's atmosphere, melt the Artic and Antarctic ice-caps and change the pattern of vegetable and animal life throughout the world, which is a frightening possibility.

VITAMINS ARE VITAL

It was not until the beginning of the twentieth century that it was recognised that certain substances were essential in the diet to prevent or cure some diseases. These substances a____ (1) known a____ (2) vitamins, a____ (3) they a____ (4) vital f____ (5) the gro____ (6), good hea____ (7), and maint____ (8) of t____ (9) normal func____ (10) of t____ (11) body. A we____ (12) balanced di____ (13) should pro____ (14) all t____ (15) vitamins w____ (16) normally req____ (17). Those o____ (18) us w____ (19) are ab____ (20) to b____ (21) sufficient fo____ (22) should n____ (23) suffer fr____ (24) vitamin defic____ (25). However, food served in restaurants and canteens has often lost much of its vitamin content because it has been kept hot, or even prepared the day before, so you may have problems if you eat out regularly.

C-Test B

NAME……...

Directions: The following test has been developed by removing the second half of every second word in four different texts beginning with the second sentence. The missing part contains the same number of letters as the first part or one more letter than the first part. No contracted forms have been used, no proper names or numbers have been deleted. You are supposed to reconstruct the texts.

Example:

A British university is now doing research into the difference between men and women drivers. It se_ _ _ that wo_ _ _ often dr_ _ _ more care_ _ _ _ _ than m_ _.

A British university is now doing research into the difference between men and women drivers. It **seems** that **women** often **drive** more **carefully** than **men**.

PHYSICAL EXERCISE

We all need exercise, especially young people in their teens and adults from twenty to eighty. Regular exercise tempo_ _ _ _ _ (1) tires t_ _ (2) body b_ _ (3) later o_ (4) actually gi_ _ _ (5) you mo_ _ (6) energy. Th_ _ (7) is w_ _ (8) people w_ _ (9) suffer fr_ _ (10) general tire_ _ _ _ _ (11) can ben_ _ _ _ (12) from tak_ _ _ (13) more exer_ _ _ _ (14) rather th_ _ (15) more re_ _ (16). Exercise ma_ _ _ (17) you fe_ _ (18) and lo_ _ (19) better a_ _ (20) can al_ _ (21) help y_ _ (22) to lo_ _ (23) weight bec_ _ _ _ (24) it bu_ _ _ (25) up fat or food to produce energy. However, if you are over 40, or if you have recently had a serious illness, you should visit your doctor before starting a general exercise routine.

RELAX AND LIVE

It is commonly believed that only rich middle-aged businessmen suffer from stress, but in fact anyone may become ill as a result of stress if they experience a lot of worry over a long period and their health is not particularly good. Stress can b____ (1) a friend o_____ (2) an en_____ (3): it c_____ (4) warn y_____ (5) that y_____ (6) are un_____ (7) too mu_____ (8) pressure a_____ (9) should cha_____ (10) your w_____ (11) of li_____ (12). It c_____ (13) kill y_____ (14) if y_____ (15) do n_____ (16) notice t_____ (17) warning sig_____ (18). Doctors ag_____(19) that i_____ (20) is prob_____ (21) the big_____ (22) cause o_____ (23) illness i_____ (24) the wes_____ (25) world. Stress can cause car accidents, heart attacks, alcoholism and may even drive people to suicide. As with all illnesses prevention is better than cure so if you are not able to relax then it is time to stop and ask yourself whether your present life really suits you.

ALTERNATIVE SOURCES OF ENERGY

The world is running out of oil and governments are searching for a suitable alternative, but so far in vain. Many scien_ _ _ _ _ (1) are optim_ _ _ _ _ (2) that n_ _ (3) ways o_ (4) generating la_ _ _ (5) amounts o_ (6) energy wi_ _ (7) be succes_ _ _ _ _ _ (8) developed, b_ _ (9) at t_ _ (10) same ti_ _ (11) they fe_ _ (12) the conseq_ _ _ _ _ _ (13). If t_ _ (14) world popul_ _ _ _ _ (15) goes o_ (16) increasing a_ (17) its pre_ _ _ _ (18) rate, a_ _ (19) each indiv_ _ _ _ _ (20) continues t_ (21) use mo_ _ (22) energy ev_ _ _ (23) year, w_ (24) will dam_ _ _ (25) the earth's atmosphere, melt the Artic and Antarctic ice-caps and change the pattern of vegetable and animal life throughout the world, which is a frightening possibility.

VITAMINS ARE VITAL

It was not until the beginning of the twentieth century that it was recognised that certain substances were essential in the diet to prevent or cure some diseases. These subst_____ (1) are kn____ (2) as vita____ (3), and th____ (4) are vi____ (5) for t_____ (6) growth, go_____ (7) health, a_____ (8) maintenance o_____ (9) the nor_____ (10) functions o_____ (11) the bo_____ (12). A well bala_____ (13) diet sho_____ (14) provide a_____ (15) the vita_____ (16) we norm_____ (17) require. Th_____ (18) of u_____ (19) who a_____ (20) able t_____ (21) buy suffi_____ (22) food sho_____ (23) not suf_____ (24) from vit_____ (25) deficiency. However, food served in restaurants and canteens has often lost much of its vitamin content because it has been kept hot, or even prepared the day before, so you may have problems if you eat out regularly.

KEYS:

KEY C-TEST A

PHYSICAL EXERCISE

We all need exercise, especially young people in their teens and adults from twenty to eighty. Regular **exercise** (1) temporally **tires** (2) the **body** (3) but **later** (4) on **actually** (5**)** gives **you** (6) more **energy** (7). This **is** (8) why **people** (9) who **suffer** (10) from **general** (11) tiredness **can** (12**)** benefit **from** (13**)** taking **more** (14**)** exercise **rather** (15) than **more** (16) rest.
Exercise (17) makes **you** (18) feel **and** (19) look **better** (20) and **can (**21) also **help (**22) you **to** (23) lose **weight** (24) because **it** (25) burns up fat or food to produce energy. However, if you are over 40, or if you have recently had a serious illness, you should visit your doctor before starting a general exercise routine.

RELAX AND LIVE

It is commonly believed that only rich middle-aged businessmen suffer from stress, but in fact anyone may become ill as a result of stress if they experience a lot of worry over a long period and their health is not particularly good. Stress **can** (1) be a **friend** (2) or **an** (3) enemy: **it** (4) can **warn** (5) you **that** (6) you **are** (7) under **too** (8) much **pressure** (9) and **should** (10**)** change **your** (11) way **of** (12) life. **It** (13) can **kill** (14) you **if** (15) you **do** (16) not **notice** (17) the **warnin**g (18) signals. **Doctors** (19) agree **that** (20) it **is** (21) probably **the** (22) biggest **cause** (23) of **illness** (24) in **the** (25**)** western word. Stress can cause car accidents, heart attacks, alcoholism and may even drive people to suicide. As with all illnesses prevention is better than cure so if you are not able to relax then it is time to stop and ask yourself whether your present life really suits you.

ALTERNATIVE SOURCES OF ENERGY

The world is running out of oil and governments are searching for a suitable alternative, but so far in vain. Many scientists **are** (1) optimistic **that** (2) new **ways** (3) of **generating** (4) large **amounts** (5) of **energy** (6) will **be** (7) successfully **developed** (8**)**, but **at** (9**)** the **same** (10**)** time **they** (11**)** fear **the** (12**)** consequences. **If** (13) the **world** (14) population **goes** (15) on **increasing** (16) at **its** (17) present **rate** (18**),** and **each** (19) individual **continue** (20) to **use** (21) more **energy** (22**)** every **year** (23), we **will** (24**)** damage **the** (25**)** earth´s atmosphere, melt the Artic and Antarctic ice-caps and change the pattern of vegetable and animal life throughout the world – a frightening possibility.

VITAMINS ARE VITAL

It was not until the beginning of the twentieth century that it was recognised that certain substances were essential in the diet to prevent or cure some diseases. These substances **are** (1) known **as** (2) vitamins, **and** (3) they **are** (4) vital **for** (5) the **growth** (6), good **health** (7) and **maintenance** (8) of **the** (9) normal **functions** (10) of **the** (11) body. A **well** (12) balanced **diet** (13) should **provide** (14) all **the** (15) vitamins **we** (16) normally **require** (17). Those **of** (18) us **who** (19) are **able** (20) to **buy** (21) sufficient **food** (22) should **not** (23) suffer **from** (24) vitamin **deficiency** (25**)**. However, food served in restaurants and canteens has often lost much of its vitamin content because it has been kept hot, or even prepared the day before, so you may have problems if you eat out regularly.

KEY C-TEST B

PHYSICAL EXERCISE

We all need exercise, especially young people in their teens and adults from twenty to eighty. Regular exercise **temporally** (1) tires **the** (2) body **but** (3) later **on** (4) actually **gives** (5) you **more** (6) energy. **This** (7) is **why** (8)people **who** (9) suffer **from** (10) general **tiredness** (11) can **benefit** (12) from **taking** (13) more **exercise** (14) rather **than** (15) more **rest** (16).
Exercise **makes** (17) you **feel** (18) and **look** (19) better **and** (20) can **also** (21) help **you** (22) to **lose** (23) weight **because** (24) it **burns** (25) up fat or food to produce energy. However, if you are over 40, or if you have recently had a serious illness, you should visit your doctor before starting a general exercise routine.

RELAX AND LIVE

It is commonly believed that only rich middle-aged businessmen suffer from stress, but infact anyone may become ill as a result of stress if they experience a lot of worry over a long period and their health is not particularly good. Stress can **be** (1) a friend **or** (2) an **enemy** (3) : it **can** (4**)** warn **you** (5) that **you** (6) are **under** (7) too **much** (8) pressure **and** (9) should **change** (10) your **way** (11) of **life** (12). It **can** (13) kill **you** (14) if **you** (15) do **not** (16) notice **the** (17) warning **signals** (18**)**. Doctors **agree** (19) that **it** (20) is **probably** (21) the **biggest** (22) cause **of** (23) illness **in** (24) the **western** (25) word. Stress can cause car accidents, heart attacks, alcoholism and may even drive people to suicide. As with all illnesses prevention is better than cure so if you are not able to relax then it is time to stop and ask yourself whether your present life really suits you.

ALTERNATIVE SOURCES OF ENERGY

The world is running out of oil and governments are searching for a suitable alternative, but so far in vain. Many **scientists** (1) are **optimistic** (2) that **new** (3) ways **of** (4) generating **large** (5) amounts **of** (6) energy **will** (7) be **successfully** (8) developed, **but** (9) at **the** (10) same **time** (11) they **fear** (12) the **consequences** (13). If **the** (14) world **population** (15) goes **on** (16) increasing **at** (17) its **present** (18) rate, **and** (19) each **individual** (20) continues **to** (21) use **more** (22) energy **every** (23) year, **we** (24) will **damage** (25) the earth´s atmosphere, melt the Artic and Antarctic ice-caps and change the pattern of vegetable and animal life throughout the world – a frightening possibility.

VITAMINS ARE VITAL

It was not until the beginning of the twentieth century that it was recognised that certain substances were essential in the diet to prevent or cure some diseases. These **substances** (1) are **known** (2) as **vitamins** (3), and **they** (4) are **vital** (5) for **the** (6) growth, **good** (7) health **and** (8) maintenance **of** (9) the **normal** (10) functions **of** (11**)** the **body** (12**)**. A well **balanced** (13) diet **should** (14) provide **all** (15**)** the **vitamins** (16) we **normally** (17) require. **Those** (18) of **us** (19) who **are** (20) able **to** (21) buy **sufficient** (22)food **should** (23) not **suffer** (24) from **vitamin** (25) deficiency. However, food served in restaurants and canteens has often lost much of its vitamin content because it has been kept hot, or even prepared the day before, so you may have problems if you eat out regularly

APÉNDICE 2

Test del Tutor o de Control

Directions: In the following sentences the last letters of some words have been removed. You are required to reconstruct the sentences.

1. I'm glad we had this opp_____ to talk.
2. There are a doz_____ eggs in the basket.
3. Every working person must pay income t_____.
4. The pirates buried the trea_____ on a desert island.
5. Her beauty and ch_____ had a powerful effect on men.
6. La_____ of rain led to a shortage of water in the city.
7. He takes cr_____ and sugar in his coffee.
8. The rich man died and left all his we_____ to his son.
9. Pu_____ must hand in their papers by the end of the week
10. This sweater is too tight. It needs to be stret_____.
11. Anne intro_____ her boyfriend to her mother.
12. Teenagers often adm_____ and worship pop singers.
13. If you blow up that balloon any more it will bu_____.
14. In order to be accepted into university, he had to impr_____ his grades.
15. The telegram was deli_____ two hours after it had been sent.
16. The differences were so sl_____ that they went unnoticed.
17. The dress you're wearing is lov_____.
18. He wasn't very popu_____ when he was a teenager, but he has many friends now

APÉNDICE 3

CUESTIONARIO

DATOS PERSONALES:

NOMBRE .....

Edad: □ Menor de 30 años □ Mayor de 30 años

Sexo: □ Varón □ Mujer

Estudios: □ Primarios □ Bachillerato □ Universitarios

Profesión.....

Años que lleva estudiando inglés:
□ menos de 3 □ entre 3-5 □ entre 5-10 □ más de 10 años

¿Utiliza usted normalmente el idioma inglés? □ Sí □ No
¿Con qué frecuencia? (sólo si la respuesta es afirmativa)
□ diariamente □ semanalmente □ mensualmente □ de vez en cuando
□ en vacaciones

¿Con qué frecuencia asiste a clase?
□ hasta el 25% □ hasta el 50% □ hasta el 75%
□ hasta el 90% □ siempre

¿Con qué frecuencia hace los deberes?
□ nunca □ hasta el 25% □ hasta el 50% □ hasta el 75%
□ siempre

¿Lee normalmente en inglés? □ Sí □ No
¿Con qué frecuencia? (sólo si la respuesta es afirmativa)
□ diariamente □ semanalmente □ mensualmente □ de vez en cuando

¿ Ve alguna vez películas en inglés? □ Sí □ No
¿Con qué frecuencia? (sólo si la respuesta es afirmativa)
□ diariamente □ semanalmente □ mensualmente □ de vez en cuando

¿Habla alguna vez con alguna persona en inglés? □ Sí □ No
¿Con qué frecuencia? (sólo si la respuesta es afirmativa)
□ diariamente □ semanalmente □ mensualmente □ de vez en cuando
□ en vacaciones

Cuándo escribe en inglés, ¿piensa normalmente en inglés o lo hace primero en su lengua materna y después lo traduce al inglés?

- Siempre pienso en mi lengua materna y después lo traduzco al inglés.
- Intento pensar en inglés pero me baso en mi lengua materna cuando encuentro una estructura o palabra que no conozco.
- Siempre pienso en inglés y cuando hay algo que no conozco lo evito o lo intento expresar de otra forma.

Cuándo hablas en inglés, ¿piensa normalmente en inglés o lo hace primero en su lengua materna y después lo traduce al inglés?

- Siempre pienso en mi lengua materna y después lo traduzco al inglés.
- Intento pensar en inglés pero me baso en mi lengua materna cuando encuentro una estructura o palabra que no conozco.
- Siempre pienso en inglés y cuando hay algo que no conozco lo evito o lo intento expresar de otra forma.

CUESTIONARIO SOBRE EL C-TEST

Queremos saber su opinión sobre la prueba que acaba de realizar. Es un nuevo tipo de examen que pretende medir su competencia global en lengua inglesa.

1 ¿Ha encontrado dificultades para realizarlo? □ Sí □ No
¿De qué tipo? (sólo si la respuesta es afirmativa)
.....
.....

2 Marque del 1 al 5 el grado en que este examen mide los distintos aspectos de la lengua. (1 nada, 2 muy poco, 3 poco, 4 bastante, 5 mucho)

aspectos gramaticales	1	2	3	4	5
ortografía: spelling	1	2	3	4	5
conocimiento general de la lengua	1	2	3	4	5
fluidez	1	2	3	4	5
léxico: vocabulario	1	2	3	4	5

3 ¿Le parece una prueba adecuada?
(1 nada, 2 poco, 3 bastante, 4 adecuada, 5 muy adecuada) 1 2 3 4 5

4 ¿Le parece un examen completo?
(1 nada, 2 poco, 3 bastante, 4 completo 5 muy completo) 1 2 3 4 5

5 ¿Cree que reflejará bien sus conocimientos de inglés?
(1 muy mal, 2 mal, 3 bastante bien, 4 bien, 5 muy bien) 1 2 3 4 5

6 ¿Le gustaría que esta prueba formara parte del ejercicio de comprensión de lectura del examen del Certificado Elemental?
□ Sí □ No

Muchas gracias por su colaboración